Elaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème de consensus en classification

Israël-César Lerman, Philippe Peter

INRIA

CENTRE DE RENNES
IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél (3) 954 90 20

Rapports de Recherche

N° 434

ÉLABORATION ET LOGICIEL D'UN INDICE DE SIMILARITÉ ENTRE OBJETS D'UN TYPE QUELCONQUE

APPLICATION AU PROBLÈME DE CONSENSUS EN CLASSIFICATION

Israël-César LERMAN
Philippe PETER

Juillet 1985

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE
Tél. : (99) 36.20.00
Télex : UNIRISA 95 0473 F

Publication Interne n° 262
Juillet 1985 - 72 pages

ELABORATION ET LOGICIEL D'UN INDICE DE SIMILARITE ENTRE OBJETS D'UN TYPE QUELCONQUE. APPLICATION AU PROBLEME DE CONSENSUS EN CLASSIFICATION

Israël-César LERMAN et
Philippe PETER

RESUME : Le problème de la définition d'un indice d'association entre variables présente moins d'ambiguité que celui d'un indice de similarité entre objets. Pour ce dernier problème, on présente une méthode très générale fine et féconde de normalisation -variable par variable- au niveau de l'ensemble des paires d'objets. Les variables descriptives sont supposées de types quelconques et on en distingue six : "numérique", "logique", "qualitatif nominal", "qualitatif ordinal", "préordonnance" et "graphe valué". L'indice peut être pris en compte par l'Algorithme de la Vraisemblance du Lien, ce qui permet d'offrir une solution significative et efficace au problème du consensus en classification.
L'aspect logiciel du calcul -dans les différents cas- est présenté conformément aux normes MODULAD, avec une description des sous-programmes et de leurs paramètres.

SUMMARY :The problem of the definition of an association coefficient between descriptive variables, is less ambiguous than those of the definition of a similarity index between objects or individuals. For this last problem we developp a very general and fruitful method which is based upon the standardization -variable by variable- at the level of the set of object pairs. We distiguich six types of descriptive variables : "numerical", "logical (0-1)", "nonimal qualitative", "ordinal qualitative", "preordonnance" and "weighted graph". The obtained similarity coefficient takes into account a mixing of the different types of descriptive variables. On the other hand, this coefficient is compatible with the Likelyhood Link Algorithm. Then it becomes possible to propose an elegant and significant and efficiency solution to the general problem of consensus in classification.
The software of the computational aspects is presented accordingly to the MODULAD standardization. Then, the subroutines and their parameters are precisely described.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

ELABORATION ET LOGICIEL D'UN INDICE DE SIMILARITE ENTRE OBETS D'UN TYPE QUELCONQUE APPLICATION AU PROBLEME DE CONSENSUS EN CLASSIFICATION

Israël-César LERMAN
Philippe PETER

Publication Interne n° 262 - Juillet 1985

PI 250 **A BCMP extension to multiserver stations with concurrent classes of customers**
Jean – Yves Le Boudec, 32 pages ; Mars 1985.

PI 251 **An approach to natural language semantics in logic programming**
Patrick Saint – Dizier, 34 pages ; Mars 1985.

PI 252 **A model to analyse the causality in synchronous real time systems**
Albert Benveniste, 28 pages ; Avril 1985.

PI 253 **Précision numérique dans le cumul d'un grand nombre de termes**
Michèle Raphalen, Bernard Philippe, 44 pages ; Avril 1985.

PI 254 **Observational congruence of non-deterministic and communicating finite processes in asynchronous systems**
Boubakar Gamatié, 22 pages ; Avril 1985.

PI 255 **Dérivation d'algorithmes distribués d'arbitrage**
Jean-Pierre Verjus, René Thoraval, 30 pages ; Mai 1985.

PI 256 **Spécification et représentation par réseaux de Pétri d'un exécutif temps réel**
Maryline Silly, Houssine Chetto, 11 pages ; Mai 1985.

PI 257 **Towards an interactive Math Mode in TEX**
Jacques André, Yann Grundt, Vincent Quint, 16 pages ; Mai 1985.

PI 258 **Experiments in teaching METAFONT**
Jacques André, Richard Southall, 16 pages ; Mai 1985.

PI 259 **Une critique de la notion de test de processus fondée sur la non séparabilité de certaines classes de langages**
Philippe Darondeau, 40 pages ; Juin 1985.

PI 260 **Contrôler les transferts de connaissance dans les algorithmes distribués - Application à la détection de l'interblocage**
Jean – Michel Hélary, Aomar Maddi, Michel Raynal, 22 pages ; Juin 1985.

PI 261 **Détecter la perte de jetons et les regénérer sur une structure en anneau**
Michel Raynal, Gérardo Rubino, 24 pages ; Juillet 1985.

PI 262 **Elaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème de consensus en classification**
Israël-César Lerman, Philippe [redacted]

ELABORATION ET LOGICIEL D'UN INDICE DE SIMILARITE ENTRE OBJETS D'UN TYPE QUELCONQUE. APPLICATION AU PROBLEME DU CONSENSUS EN CLASSIFICATION

I.C. LERMAN et Ph. PETER

I. INTRODUCTION

Notre but consiste ici à montrer une solution pour le problème délicat de la définition d'un indice de similarité entre éléments d'un ensemble E d'objets (n=card(E)) décrit par un ensemble V de variables (m=card(V)) de types quelconques.

Cette solution repose sur une technique très générale, précise et féconde, de normalisation -variable par variable- au niveau de l'ensemble $P_2(E)$ des paires d'objets distincts de E. Une telle méthode tient finement compte du caractère relatif de la ressemblance entre deux objets, eu égard à l'ensemble des objets où ils se situent.

Nous commencerons par supposer -ce qui est fréquemment le cas- que toutes les variables sont d'un même type. A cet égard, nous disposons d'une claire typologie des variables de description (cf. (LERMAN(1981)) Chap.2) et nous distinguerons ici six types ou cas : "numérique", "logique (0-1)", "qualitatif nominal", "qualitatif ordinal", "préordonnance" et "graphe valué". Ces différents cas de figure d'un tableau de données seront ci-dessous explicités. Enfin, et de façon très particulière, nous considérons le cas d'un tableau de contingence puis celui de la juxtaposition de tels tableaux.

Nous avons suffisamment insisté dans nos précédentes parutions (cf. par exemple la référence ci-dessus mentionnée) qu'en dehors de la table de contingence, un tableau de données Objets x Variables est de nature mathématique essentiellement dissymétrique : Variables et Objets ne jouent pas vis à vis les uns des autres le même rôle et il suffit pour s'en convaincre d'examiner le cas où les variables sont qualitatives.

Une bonne partie de notre recherche a porté sur la comparaison de variables d'un même type. D'une certaine façon -liée à ce que nous venons d'exprimer dans le dernier alinéa- ce dernier problème présente moins d'ambiguité que celui qui nous occupe ici. En effet, dans l'évaluation de l'association entre deux variables, les différents objets de l'échantillon qui définit E ont exactement la même importance et ont a priori à intervenir de façon égale. Alors que dans notre problème -une fois normalisées les contributions des différentes variables- on ne sait pas s'il n'y a pas lieu d'accorder une plus forte pondération à certaines variables. Si on se conforme à l'opinion des promoteurs de la "Taxonomie Numérique" ((SNEATH & SOKAL (1972))), c'est avec une égale importance que les variables doivent intervenir pour contribuer à la ressemblance entre deux objets. De toute façon, notre procédure de normalisation -variable par variable- permettra de façon très souple de prendre en compte une pondération des variables, posée a priori ou résultant d'une technique objective (e.g. analyse factorielle ou "importance projective").

Notre méthode permettra aisément de prendre en compte un ensemble V de variables de description dont les types diffèrent. En cela, nous répondons au même problème que celui de J.C. Gower (GOWER(1971)) dont le coefficient n'intègre pas des échelles de variables qualitatives aussi fines et riches que les nôtres et d'autre part, n'obéit pas à notre critère de normalisation.

Terminons cette introduction en signalant que ce texte reprend et enrichit sensiblement le paragraphe VI du chapitre 2 de (LERMAN(1981)). En effet, on intègre les variables de types "préordonnance" et "graphe valué". D'autre part, on remplace une normalisation globale par celle -plus fine- considérée ici, variable par variable.

II. CAS OU LES VARIABLES SONT NUMERIQUES

Désignons par $\{o_i/i\in I\}$ -où $I=\{1,2,\ldots,i,\ldots,n\}$- l'ensemble E des objets et par $\{v^j/j\in J\}$ -où $J=\{1,2,\ldots,j,\ldots,m\}$- l'ensemble V des variables qu'on suppose ici numériques et à valeurs positives. Cette restriction de positivité est en réalité tout à fait mineure puisque de façon quasi-générale, les variables numériques qui se présentent en analyse des données sont naturellement à valeurs positives.

L'appréhension de l'objet o_i se fait à partir de la suite des mesures $(x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^m)$, où $x_i^j = v^j(o_i)$ est la mesure de la j-ème variable sur o_i.

La comparaison de deux objets o_i et $o_{i'}$ pris dans E a déjà conduit aux trois indices suivants qui ont été expérimentés dans le cadre de l'"Algorithme de la Vraisemblance du Lien" :

$$COR(o_i, o_{i'}) = \frac{\sum_j (x_i^j - x_i^\bullet)(x_{i'}^j - x_{i'}^\bullet)}{\sqrt{\sum_j (x_i^j - x_i)^2 \sum_j (x_{i'}^j - x_{i'})^2}}, \quad (1)$$

où $x_i^\bullet$ est la moyenne des mesures des différentes variables sur l'objet o_i,

$$COS(o_i, o_{i'}) = \frac{\sum_j (x_i^j x_{i'}^j)}{\sqrt{\sum_j (x_i^j)^2 \sum_j (x_{i'}^j)^2}}, \quad (2)$$

et

$$L(o_i, o_{i'}) = \sum_j \frac{(x_i^j - x_\bullet^j)(x_{i'}^j - x_\bullet^j)}{s^2(j)}, \quad (3)$$

où $s^2(j) = \frac{1}{n} \sum_i (x_i^j - x_\bullet^j)^2$ est la variance empirique de la variable v_j.

L'indice (1) est celui bien connu du coefficient de corrélation, mais entre objets. Il reste largement utilisé par les Taxonomistes numériciens. Il possède la propriété intéressante d'invariance lorsqu'on remplace l'un ou l'autre des deux objets à associer par son "homothétique" où la suite des mesures des différentes variables se trouvent multipliées par le même coefficient. Néanmoins, l'interprétation de la moyenne $x_i^\bullet$ (resp. $x_{i'}^\bullet$) n'est pas claire et ce d'autant plus que les échelles des valeurs prises par les différentes variables sont hétérogènes (du point de vue de leurs amplitudes respectives et de leurs variances respectives).

La propriété d'invariance que nous venons de mentionner est préservée dans l'indice (2) d'expression plus simple, d'interprétation plus claire et qui est défini -en termes géométriques- par le cosinus de l'angle des deux vecteurs de R^m dont les composantes respectives sont $(x_i^1,\ldots,x_i^j,\ldots,x_i^m)$ et $(x_{i'}^1,\ldots,x_{i'}^j,\ldots,x_{i'}^m)$. D'ailleurs, à partir des expériences menées sur un grand ensemble de jeux de données, c'est l'indice (2) que nous retenons préférentiellement à (1).

Toutefois, dans le cas où il existe un grand degré d'hétérogénéité entre les variables (du point de vue de leurs amplitudes et variances respectives), on peut à première vue considérer l'indice (3), d'ailleurs sous-jacent à l'Analyse en Composantes Principales Normée et qui correspond au produit scalaire des vecteurs des mesures centrées réduites, variable par variable : $(\xi_i^1,\ldots,\xi_i^j,\ldots,\xi_i^m)$ et $(\xi_{i'}^1,\ldots,\xi_{i'}^j,\ldots,\xi_{i'}^m)$, où $\xi_i^j=(x_i^j-x_\cdot^j)/s_j$ (resp. $\xi_{i'}^j=(x_{i'}^j-x_\cdot^j)/s_j$).

C'est le cosinus de l'angle entre ces deux derniers vecteurs qui a permis à J.R. Massé (MASSE(1978)) d'obtenir les résultats les plus significatifs dans un problème de classification de composants électroniques d'origines diverses, sur lesquels ont été effectuées différentes mesures électriques où d'une sous-classe de variables à une autre, l'ordre de grandeur de la variance pouvait être multiplié par 10, par 100 ou même par 1000.

Toutefois, dans de nombreuses études où les rapports entre les variances des différentes variables étaient relativement importants, un indice tel que (3) a donné des résultats plus parcellaires que ceux (1) et (2). En effet, ce dernier indice ne possède plus la propriété d'invariance par homothétie. De façon liée, la mesure d'une variable se trouve déconnectée de la mesure des autres variables sur le même objet.

Pourtant, on peut se poser la question de savoir pourquoi la variance empirique d'une variable descriptive doit pondérer son importance (par rapport aux autres variables) pour l'évaluation de la ressemblance entre deux objets.

L'indice que nous allons proposer commence par relativiser la mesure d'une même variable par rapport à celles des autres sur un même objet et en cela, il possède la propriété d'invariance par homothétie. Désignons par η_{ij} cette mesure relative de la variable v^j sur l'objet o_i, $1 \leq i \leq n$, $1 \leq j \leq m$.

Pour une même variable v^j, en ce qui concerne la comparaison de deux objets o_i et $o_{i'}$, on commence par admettre une forme multiplicative de l'indice brut de ressemblance ($s_j(i,i') = \eta_{ij}\eta_{i'j}$). La contribution de la variable v^j pour l'évaluation de la ressemblance entre i et i', résulte alors de la normalisation sur l'ensemble $P_2(I)$ des paires d'éléments de I, de $s_j(i,i')$.

Une telle normalisation rend bien compte du caractère relatif de la ressemblance entre deux objets par rapport à l'ensemble des paires d'objets de l'ensemble où ils se situent et qui est à organiser. D'autre part, la relative hétérogénéité des différentes variables se trouve neutralisée lorsqu'on tient compte -de façon additive- de l'ensemble des variables pour l'évaluation de la ressemblance entre deux objets. Soulignons le fait que cette normalisation s'effectue directement au niveau de $P_2(I)$ et non au niveau de I comme c'est le cas pour l'indice (3).

Dans (LERMAN(1981), Chap.2 § VI) nous avions bien proposé ce type de normalisation, mais pour un indice directement global tel que (2). Ce qu'il y a de différent ici, c'est que la normalisation doive d'abord être effectuée variable par variable, de façon que l'indice se présente comme une somme de contributions normalisées. Pour ce dernier, on considère une nouvelle normalisation de même type avant l'application de l'Algorithme de la Vraisemblance du Lien.

Cette technique qui donne des résultats raffinés a -comme nous l'avons déjà mentionné- l'intérêt de se généraliser à n'importe quels types de variables.

1. Contribution brute d'une variable à la comparaison de deux objets

Pour ne pas déconnecter la mesure d'une variable v^j des mesures des autres variables sur le même objet o_i dont il y a lieu de préserver l'entité, nous proposerons comme mesure réduite de la j-ème variable sur o_i :

$$w^j(o_i) = \frac{v^j(o_i)}{\sqrt{\sum_j (v^j(o_i))^2}} = \frac{x_i^j}{\sqrt{\sum_j (x_i^j)^2}} = \eta_i^j. \quad (4)$$

Une telle mesure réduite est évidemment invariante par homothétie portée sur la suite des mesures initiales.

La contribution brute de la j-ème variable à la comparaison de deux objets o_i et $o_{i'}$ sera de façon multiplicative posée comme suit :

$$s_j(o_i,o_{i'}) = \eta_i^j \eta_{i'}^j, \quad . \quad (5)$$

La somme pour j=1,...,m, de $s_j(o_i,o_{i'})$ est l'indice $COS(o_i,o_{i'})$ (cf.(2)) que nous ne considérons pas ici, mais dont nous nous sommes inspirés pour déterminer les contributions élémentaires (4) et (5).

2. Moyenne sur $P_2(I)$ de $s_j(o_i,o_{i'})$

Nous désignerons par M^j cette moyenne. On a :

$$M^j = \frac{2}{n(n-1)} \sum\{\eta_i^j \eta_{i'}^j, / \{i,i'\} \in P_2(I)\} \quad (6)$$

$$= \frac{n}{(n-1)} \{(\mu(w^j))^2 - \frac{1}{n} \mu_2(w^j)\}, \quad (7)$$

où $\mu(w^j)$ et $\mu_2(w^j)$ sont la moyenne et le moment d'ordre 2 de w^j :

$$\mu(w^j) = \eta_\cdot^j = \frac{1}{n} \sum_i \eta_i^j$$

et

$$\mu_2(w^j) = \frac{1}{n} \sum_i (\eta_i^j)^2$$

Le passage de (6) à (7) repose sur l'identité

$$2\sum\{\eta_i^j \eta_{i'}^j, / \{i,i'\} \in P_2(I)\}$$

$$= (\sum_i \eta_i^j)^2 - \sum_i (\eta_i^j)^2$$

3. Variance sur $P_2(I)$ de $s_j(o_i,o_{i'})$

La structure du calcul du moment absolu d'ordre 2 est la même que celle de la moyenne M^j ci-dessus. On a -en désignant par M_2^j ce moment-

$$M_2^j = \frac{n}{(n-1)}\left\{\left[\mu_2(w^j)\right]^2 - \frac{1}{n}\mu_4(w^j)\right\}, \quad (8)$$

où $\mu_4(w^j) = \frac{1}{n}\sum_i (\eta_i^j)^4$ est le moment absolu d'ordre 4 de w^j.

On a, très sensiblement,

$$M^j = \left\{(\mu(w^j))^2 - \frac{1}{n}\mu_2(w^j)\right\} \quad (9)$$

et

$$M_2^j = \left\{(\mu_2(w^j))^2 - \frac{1}{n}\mu_4(w^j)\right\}. \quad (10)$$

De sorte que la variance $(\sigma^j)^2$ s'écrit

$$(\sigma^j)^2 = (\mu_2(w^j))^2 - (\mu(w^j))^4$$

$$- \frac{1}{n}\left\{\mu_4(w^j) - 2\mu_2(w^j)(\mu(w^j))^2 + \frac{1}{n}(\mu_2(w^j))\right\}, \quad (11)$$

dont la partie dominante est

$$(\mu_2(w^j))^2 - (\mu(w^j))^4. \quad (12)$$

4. L'indice

L'indice que nous proposons entre les deux objets o_i et $o_{i'}$ se met sous la forme de la somme réduite des contributions normalisées des différentes variables :

$$S(o_i,o_{i'}) = \frac{1}{\sqrt{m}} \sum_{1\leq j\leq m} (s_j(o_i,o_{i'}) - M^j)/\sigma^j, \quad (13)$$

où la réduction au moyen de $1/\sqrt{m}$ se réfère à un modèle d'indépendance où les variables aléatoires associées aux v^j, $1\leq j\leq m$, ont une variance unité.

De toute façon, cette réduction n'intervient plus après la réduction globale des similarités où on substitue à la table

$$\{S(o_i,o_{i'})/\{i,i'\} \in P_2(I)\}, \quad (14)$$

celle

$$\{T(o_i,o_{i'})/\{i,i'\} \in P_2(I)\}, \quad (15)$$

avec

$$T(o_i,o_{i'})=(S(o_i,o_{i'})-\bar{S})/\sqrt{var(S)}, \quad (16)$$

où $\bar{S}$ et var(S) sont respectivement la moyenne et la variance de la table (14).

La table qui est directement l'argument de l'algorithme de la vraisemblance du lien, se met sous la forme

$$\{P(i,i')/\{i,i'\} \in P_2(I)\}, \quad (17)$$

où

$$P(i,i') = \Phi(T(o_i,o_{i'})),$$

où Φ est la fonction de répartition de la loi N(0,1), normale centrée-réduite.

5. Prise en compte d'une pondération des variables

Nous avons déjà signalé que la prise en compte a priori d'une pondération des variables est tout à fait problématique et discutable, même si cette pondération est basée sur une méthode objective ([LERMAN(1970b)], [SNEATH & SOKAL(1972)]). Néanmoins, il faut laisser la porte ouverte aux possibilités expérimentales de l'analyse classificatoire des données.

Si $\alpha_1,\alpha_2,\ldots,\alpha_j,\ldots,\alpha_m$ est une suite de coefficients positifs de somme unité, définissant les importances respectives qu'on veut donner aux différentes variables pour l'évaluation de la similarité entre deux objets. La forme additive de l'indice S(cf.(13)) permet de les intégrer au moyen de l'expression suivante :

$$S'(i,i';\alpha_1,\ldots,\alpha_m)=\sqrt{m}\sum_{1\leq j\leq m}\alpha_j(s_j(i,i')-M^j)/\sigma^j, \quad (18)$$

où nous avons noté i pour o_i.

Certains spécialistes de l'analyse factorielle conseillent la classification des objets décrits par les quelques premiers facteurs. Dans cette démarche, c'est une pondération implicite des variables qui est -par rapport au modèle factoriel- prise en compte. Cette pondération peut être plus explicitement mise en évidence dans notre indice si on lie α_j au coefficient défini par la somme des carrés des coefficients de corrélation entre v^j et les quelques facteurs retenus, $1 \leq j \leq m$.

On peut envisager d'autres pondérations objectives. Nous avons pu définir -de façon intrinsèque au tableau des données- "l'importance projective" d'une même variable au moyen de la variance de ses indices d'association avec les autres variables (LERMAN(1981),Chap.3). Dans ces conditions, on peut lier α_j à "l'importance projective" de la variable v^j, $1 \leq j \leq m$.

Nous nous sommes ci-dessus posé la question de savoir dans quelle mesure la variance d'une variable doit discriminer la ressemblance entre deux objets. Dans notre démarche ci-dessus, nous avons apporté une forme de neutralisation de cette variance, mais au niveau de l'ensemble des paires d'objets. On peut -compte tenu de la forme de l'indice- chercher à réintroduire l'importance relative des différentes variances s_j^2 (s_j^2=var.(v^j)), $1 \leq j \leq m$, à partir de coefficients : $\alpha_1, \alpha_2, \ldots, \alpha_j, \ldots, \alpha_m$, en liant α_j à s_j^2 d'une façon qui reste à préciser.

Ce problème de pondération se pose dans les mêmes termes quel que soit le type de variable. De sorte que dans la suite nous n'évoquerons plus ce problème. Toutefois, en ce qui concerne les méthodes "objectives" de pondération, certaines notions de variance et d'analyse factorielle deviennent -dans le cas où les variables sont qualitatives ou relationnelles- très délicates à circonscrir.

III. CAS OU LES VARIABLES SONT DES ATTRIBUTS DE DESCRIPTION (VARIABLES LOGIQUES (0-1)).

Nous allons suivre les mêmes étapes que dans le cas où les variables sont numériques.

Si on ignore la normalisation -variable par variable- on peut a priori se référer à trois indices d'inspirations voisines. Le premier a la même

expression formelle que le cosinus (cf. formule (2) § II) et correspond à l'indice d'Ochiai (OCHIAI(1957)). Il peut se mettre sous la forme:

$$S_o(i,i') = \frac{\sum_j \xi_i^j \xi_{i'}^j}{\sqrt{(\sum_j \xi_i^j)(\sum_j \xi_{i'}^j)}}, \quad (1)$$

où ξ_i^j (resp.$\xi_{i'}^j$)=1 ou 0 selon que l'attribut j est présent ou absent chez l'objet o_i (resp.$o_{i'}$).

Les deux autres indices résultent directement -par transposition des rôles des lignes et des colonnes- de ceux conçus pour la comparaison des attributs de description (LERMAN(1981), Chap.2). A l'exception de la structure parfaitement symétrique d'un tableau de contingence et de ses dérivés, on peut davantage admettre -que dans le cas des autres types de tableaux de données- cette transposition des rôles.

Ces deux indices peuvent respectivement être mis sous la forme :

$$S_1(i,i') = \frac{\sum_j (\xi_i^j - p_i^\cdot)(\xi_{i'}^j - p_{i'}^\cdot)}{m\sqrt{p_i^\cdot(1-p_i^\cdot)p_{i'}^\cdot(1-p_{i'}^\cdot)}}, \quad (2)$$

où $p_i^\cdot$(resp. $p_{i'}^\cdot$) est la proportion d'attributs présents chez l'objet i (resp. i') et

$$S_2(i.i') = \frac{\sum_j (\xi_i^j - p_i^\cdot)(\xi_{i'}^j - p_{i'}^\cdot)}{m\sqrt{p_i^\cdot p_{i'}^\cdot}}. \quad (3)$$

Lorsque les indices (2) et (3) sont conçus dans la situation transposée de la comparaison d'attributs de description, c'est une même méthode (cf. référence ci-dessus mentionnée) qui nous permet de les obtenir, respectivement par rapport à deux modèles aléatoires de l'hypothèse d'absence de liaison. L'indice (2) correspond à celui de K. Pearson.

1. <u>Contribution brute d'un attribut a^j à la comparaison de deux objets o_i et $o_{i'}$.</u>

Nous allons considérer a priori trois formes de cette contribution qui, respectivement, correspondent à chacun des indices (1),(2) et (3) :

$$s_{0j}(i,i') = \frac{\xi_i^j \xi_{i'}^j}{\sqrt{p_i^\cdot p_{i'}^\cdot}}, \quad (4)$$

$$s_{1j}(i,i') = \frac{(\xi_i^j - p_i^\cdot)(\xi_{i'}^j - p_{i'}^\cdot)}{\sqrt{p_i^\cdot(1-p_i^\cdot)p_{i'}^\cdot(1-p_{i'}^\cdot)}}, \quad (5)$$

et

$$s_{2j}(i,i') = \frac{(\xi_i^j - p_i^\cdot)(\xi_{i'}^j - p_{i'}^\cdot)}{\sqrt{p_i^\cdot p_{i'}^\cdot}}. \quad (6)$$

En posant

$$\eta_{0i}^j = \frac{\xi_i^j}{\sqrt{p_i^\cdot}}, \quad (4')$$

$$\eta_{1i}^j = \frac{(\xi_i^j - p_i^\cdot)}{\sqrt{p_i^\cdot(1-p_i^\cdot)}} \quad (5')$$

et

$$\eta_{2i}^j = \frac{(\xi_i^j - p_i^\cdot)}{\sqrt{p_i^\cdot}}, \quad (6')$$

on a

$$s_{0j}(i,i') = \eta_{0i}^j \eta_{0i'}^j, \quad (4'')$$

$$s_{1j}(i,i') = \eta_{1i}^j \eta_{1i'}^j, \quad (5'')$$

et

$$s_{2j}(i,i') = \eta_{2i}^j \eta_{2i'}^j. \quad (6'').$$

2. Moyenne et Variance sur $P_2(I)$ de $s_{\alpha j}(i,i')$;
Proposition de l'indice de similarité.
(α=0,1 ou 2 conformément aux expressions (4"),(5") et (6")).

Les calculs sont ceux des paragraphes II.2 et 3 ci-dessus. On remplacera selon les cas n_i^j par n_{0i}^j, n_{1i}^j ou n_{2i}^j. En désignant par M_α^j et $(\sigma_\alpha^j)^2$ la moyenne et la variance -sur $P_2(I)$- de $s_{\alpha j}(i,i')$, on a l'expression de l'indice qui correspond à celui (13) du paragraphe II ci-dessus :

$$S_\alpha(o_i,o_{i'}) = \frac{1}{\sqrt{m}} \sum_{1<j<m} (s_{\alpha j}(o_i,o_{i'}) - M_\alpha^j)/\sigma_\alpha^j, \quad (7)$$

lequel pourra prendre trois formes selon que α=0,1 ou 2.

Les dernières étapes avant l'application de l'algorithme de la vraisemblance du lien correspondent à la réduction globale -sur $P_2(I)$- des similarités S et à la transformation de l'échelle de mesure des similarités en une échelle de probabilité ou de fréquence mathématique. Les expressions formulées sont celles (14) à (18) du paragraphe II où il y a lieu de substituer S_α à S, T_α à T et P_α à P, avec α=0,1 ou 2 correspondants aux trois formes de l'indice.

IV. CAS OU LES VARIABLES SONT QUALITATIVES NOMINALES

Désignons par C l'ensemble des variables qui sont ici des caractères descriptifs où l'ensemble des modalités d'un même caractère n'est muni d'aucune structure. On désigne ici -et dans la suite- par Q le cardinal de C. J_q indiquera l'ensemble des codes des modalités de la variable qualitative c_q ($c_q \in C$, $1<q<Q$). De façon plus précise, on posera $J_q=\{j_q / 1<j_q<m_q\}$ où m_q est le nombre de modalités de la variable c_q, $1<q<Q$.

On se ramène à un codage en "0-1" du type "absence-présence" en associant à chacune des modalités d'un même caractère un attribut de description que nous appelons "attribut-modalité" et dont la valeur est 0 ou 1 selon que la modalité en question n'est pas ou est possédée. On a ainsi ce que les factorialistes appellent un "codage disjonctif complet". Ainsi, le codage de la réponse d'un individu ou objet au caractère c_q à m_q modalités est un vecteur logique à m_q composantes dont la j_q-ème est égale à 1 et les autres à 0, si et seulement si l'individu ou objet possède la j_q-ème modalité du caractère c_q.

De la sorte, la représentation d'un objet -par rapport à la suite $\{c_q/1\leq q\leq Q\}$ des variables- se fait au moyen d'un vecteur logique à $m=\sum\{m_q/1\leq q\leq Q\}$ composantes dont exactement Q sont égales à 1 et où la q-ème composante égale 1 se situe entre la $(m_1+m_2+\ldots+m_{(q-1)}$ position et celle $(m_1+m_2+\ldots+m_q)$. Cependant, on aura soin d'effectuer tout calcul à partir du codage initial qui est beaucoup plus économique en place mémoire.

1. Contribution brute d'une variable c_q à la comparaison de deux objets o_i et $o_{i'}$

Nous sommes dans une situation où le nombre de composantes égales à 1 dans la description de chaque objet est constante et égale à Q. Il est dans ces conditions important de remarquer qu'il n'y a pas lieu de réduire la contribution de la "mesure" d'une même variable c_q par rapport à l'ensemble de toutes les variables.

De façon tout à fait naturelle, on posera

$$s_c(i,i')=\begin{cases}1 \text{ si } o_i \text{ et } o_{i'} \text{ possèdent la même modalité de c}\\ 0 \text{ si } o_i \text{ et } o_{i'} \text{ ne possèdent pas la même modalité de c.}\end{cases}$$

Dans ces conditions, en désignant par $I_1^c, I_2^c, \ldots, I_{m_q}^c$ la partition de I définie par la variable qualitative c_q, on a :

$$s_c(i,i')=1 \Leftrightarrow \{i,i'\} \in \sum_{1\leq j_q\leq m_q} P_2(I_{j_q}), \quad (1)$$

où, rappelons-le, $P_2(I_{j_q})$ est l'ensemble des paires ou parties à deux éléments de I_{j_q}.

Nous allons à présent calculer la moyenne et la variance empiriques de $s_c(i,i')$ sur l'ensemble $P_2(I)$. $s_c(i,i')$ définit la contribution brute de c à la comparaison de o_i et $o_{i'}$. On désignera par n_j^c le cardinal de I_j^c, $1\leq j\leq m_q$.

2. Moyenne et variance de $s_c(i,i')$. Proposition de l'indice de similarité.

On a, en vertu de (1)

$$\sum\{s_c(i,i')/\{i,i'\} \in P_2(I)\} = \sum_{1\leq j\leq m_q} (n_j^c(n_j^c-1)/2), \quad (2)$$

de sorte que la moyenne de $s_c(i,i')$ -sur $P_2(I)$- se met sous la forme :

$$M^c = \sum_{1\leq j\leq m_q} (n_j^c(n_j^c-1)/n(n-1)). \quad (3)$$

On a

$$s_c^2(i,i') = s_c(i,i'),$$

de sorte que le moment d'ordre 2 de $s_c(i,i')$ s'écrit

$$M_2^c = \sum_{1\leq j\leq m_q} (n_j^c(n_j^c-1)/n(n-1)). \quad (4)$$

Finalement,

$$(\sigma^2)^c = \text{var.}(s_c(i,i')) = \sum_{1\leq j\leq m_c} (n_j^c(n_j^c-1)/n(n-1)) - \sum_{1\leq j\leq m_c} (n_j^c(n_j^c-1)/n(n-1)))^2. \quad (5)$$

Dans ces conditions, l'indice de similarité entre les deux objets o_i et $o_{i'}$, tenant également compte de l'ensemble des variables, se met sous la forme :

$$S(o_i,o_{i'}) = \frac{1}{\sqrt{Q}} \sum_{1\leq q\leq Q} \frac{s_{c_q}(i,i')-M^{c_q}}{\sqrt{\sigma^{c_q}}}, \quad (6)$$

avec des notations ci-dessus explicitées.

On reprendra ici le sens de l'expression du dernier alinéa du paragraphe III ci-dessus.

V. CAS OU LES VARIABLES SONT QUALITATIVES ORDINALES

Les notations sont exactement les mêmes qu'au paragraphe IV ci-dessus. La différence est que l'ensemble J_q des modalités d'une même variable c_q, se trouve muni d'un ordre total pour lequel le rang de la j_q-ème modalité est j_q, $1 \leq j_q \leq m_q$. On posera

$c_q(o_i)=(j_q-1)$ si l'objet o_i possède la j_q-ème modalité de la variable c_q.

Nous avons été conduits dans (LERMAN(1981), Chap.2, §VI.3.) à prendre comme contribution brute d'une variable qualitative ordinale c_q à la comparaison de deux objets o_i et $o_{i'}$, l'expression

$$s_{c_q}(i,i')=(m_{c_q}-1)-|c_q(i)-c_q(i')| \qquad (1)$$

Cette expression fait suite à un codage de la réponse à la variable c_q au moyen d'un vecteur logique à $2(m_q-1)$ composantes où l'attribut défini par la h-ème s'exprime de la façon suivante :

"code initial j_q strictement inférieur à (m_q-h+1)" pour $1 \leq h < m_q$,
"code initial j_q supérieur ou égal à $(2m_q-h)$" pour $h \geq m_q$.

De la sorte, il y a exactement (m_q-1) composantes égales à 1 qui sont réparties aux extrémités du vecteur logique codant la réponse de l'objet o_i et où $c_q(o_i)$ est exactement le nombre de composantes égales à 1 qui se trouvent à l'extrémité droite du vecteur.

Ainsi, relativement à l'ensemble C des variables. le vecteur logique de description d'un même objet comporte un nombre de composantes égales à 1, indépendant de l'objet et égal $\sum\{(m_q-1)/1 \leq q \leq Q\}$. Dans ces conditions -comme dans le cas nominal- il n'y a pas lieu de réduire la contribution de la "mesure" d'une même variable c_q par rapport à l'ensemble de toutes les variables.

1. Moyenne et variance de $s_c(i,i')$. Proposition de l'indice de similarité

Le calcul repose sur la décomposition de $P_2(I)$ conformément à la partition $\{I_j^c / 1 \leqslant j \leqslant m\}$. On a

$$P_2(I) = \sum_{1 \leqslant j \leqslant m} P_2(I_j^c) + \sum_{1 \leqslant j < h \leqslant m} I_j^c * I_h^c, \quad (2)$$

où $I_j^c * I_h^c$ désigne l'ensemble des paires $\{i,i'\}$ où $i \in I_j^c$ et $i' \in I_h^c$.

Le calcul de la moyenne de $s_c(i,i')$ sur $P_2(I)$ repose sur celui de la somme de $|c(i)-c(i')|$. Dans cette dernière :

$$\sum \{|c(i)-c(i')| / \{i,i'\} \in P_2(I)\}, \quad (3)$$

la contribution de $\{i,i'\}$ est nulle si $\{i,i'\} \in P_2(I_j^c)$, $1 \leqslant j \leqslant m$. Dans ces conditions, la somme (3) se réduit à

$$\sum_{1 \leqslant j < h \leqslant m} \; \sum_{\{i,i'\} \in I_j^c * I_h^c} (|c(i)-c(i')|)$$

$$= \sum_{1 \leqslant j < h \leqslant m} n_j^c n_h^c (h-j). \quad (4)$$

D'où, la moyenne de $s_c(i,i')$:

$$M^c = (m-1) - \sum_{1 \leqslant j < h \leqslant m} \frac{2\, n_j^c n_h^c}{n(n-1)} \times (h-j) \quad (5)$$

et le moment absolu d'ordre 2 de $|c(i)-c(i')|$:

$$\sum_{1 \leqslant j < h \leqslant m} \frac{2\, n_j^c n_h^c}{n(n-1)} \times (h-j)^2 \quad (6)$$

D'où la variance $(\sigma^2)^c$ de $s_c(i,i')$:

$$V^c = \sum_{1 \leqslant j < h \leqslant m} \frac{2\, n_j^c n_h^c}{n(n-1)} \times (h-j)^2 - \left[\sum_{1 \leqslant j < h \leqslant m} \frac{2\, n_j^c n_h^c}{n(n-1)} \times (h-j) \right]^2. \quad (7)$$

Comme dans le cas qualitatif nominal, l'indice de similarité entre deux objets o_i et $o_{i'}$, tenant également compte de l'ensemble des variables, se met sous la forme :

$$S(o_i,o_{i'})=\frac{1}{\sqrt{Q}}\sum_{1\leq q\leq Q}\frac{s_{c_q}(i,i')-M^{c_q}}{\sqrt{\sigma^{c_q}}}. \quad (8)$$

On continuera à reprendre ici le sens de l'expression du dernier alinéa du paragraphe III.

VI. CAS OU LES VARIABLES SONT DES PREORDONNANCES OU DES GRAPHES VALUES

Les notations sont les mêmes qu'aux paragraphes IV et V ci-dessus.

VI.1. Structure de similarité sur J_q (q fixé)

Tout en restant extrêmement générale, la structure descriptive la plus riche et la moins arbitraire d'une variable qualitative est fournie par une préordonnance totale sur l'ensemble de ses modalités (cf. aussi. dans un tout autre contexte. (CHAH(1984))).

Dans ces conditions -relativement à la variable c_q- on introduit l'ensemble suivant des couples de modalités :

$$H_q=\{(j_q,h_q)/1\leq j_q\leq h_q\leq m_q\}, \quad (1)$$

sur lequel se trouve défini -par l'expert- un préordre total ω_q (i.e. préordonnance sur J_q) pour lequel un couple (j_q,h_q) est d'autant plus grand -d'un point de vue ordinal- que la modalité j_q ressemble à celle h_q ; ainsi, la dernière classe de ce préordre comporte m_q termes de la forme $(j_q,j_q),1\leq j_q\leq m_q$.

Exemple : Dans le cas d'un problème d'organisation d'un important corpus de petites annonces immobilières, on considère la variable "objet de la transaction" dont la suite des modalités -respectivement codées 1,2,3,4,5,6,7,8,9- est : "maison", "pavillon", "appartement", "habitation", "studio", "chambre", "local", "garage", "terrain". On peut proposer la préordonnance :

15∿16∿17∿18∿19∿25∿26∿27∿28∿29∿36∿37∿38∿39∿46∿47∿48∿49∿57∿58∿59∿67∿68∿69
78∿79∿89⟨13∿23∿35∿45⟨14∿24∿34∿56⟨12⟨11∿22∿33∿44∿55∿66∿77∿88∿99, où ij avec
1≤j indique le couple (i,j).

Dans l'introduction même de H_q, nous admettons le caractère symétrique de la notion de similarité. Sinon, comme cela pourrait se présenter dans un problème d'affectation, il suffit de définir le préordre total sur l'ensemble $K_q=J_q xJ_q$ et les mêmes considérations ci-dessous -conceptuelles et de calcul- restent valables.

A chaque élément de l'ensemble préordonné (H_q dans notre cas) on associe un "rang". Pour définir précisément la fonction ordinale "rang" désignons par $(\ell_1,\ell_2,\ldots,\ell_k)$ la suite des cardinaux de la suite ordonnée des classes du préordre total. Le rang d'un élément appartenant à la j-ème classe, $1\leq j\leq k$, est posé égal à

$$\sum_{1\leq i\leq (j-1)} \ell_i+(\ell_j+1)/2$$

Ainsi, relativement à l'exemple ci-dessus, le rang de l'élément 24 est égal à 27+4+2=33. De la sorte, la somme de tous les rangs est -comme dans le cas totalement et strictement ordinal- égal à $L(L+1)/2$, où $L=\ell_1+\ell_2+\ldots+\ell_k$. La structure descriptive sera donc basée sur le tableau des rangs ainsi calculés :

$$\{r_{j_q h_q}/(i_q,h_q) \in H_q\}. \quad (2)$$

Une autre forme, plus riche mais moins générale de la relation de similarité sur J_q, pour une fine description de E, est fournie au moyen d'une table numérique indexée par $K_q=J_q xJ_q$, où le nombre qui se trouve à l'intersection de la ligne j_q et de la colonne h_q est sensé "mesurer" le degré de ressemblance entre les deux modalités j_q et h_q. Cette table de nombres qu'on peut admettre -sans que cela soit nécessaire pour les calculs- symétrique, est supposée donnée par l'"expert". Nous l'écrivons sous la forme

$$\{p_{j_q h_q}/(j_q,h_q) \in J_q xJ_q\},$$

ou, plus simplement, en tenant compte de la symétrie,

$$\{p_{j_q h_q}/(j_q,h_q) \in H_q\}. \quad (3)$$

En réalité, la nature des calculs sera exactement la même qu'on travaille avec la table (2) des rangs, ou avec celle (3) des coefficients numériques, de sorte qu'on désignera par

$$\{s_{j_q h_q} / (j_q, h_q) \in H_q\}, \quad (4)$$

l'une ou l'autre des deux tables.

Si l'objet o_i(resp.$o_{i'}$) possède la modalité j_q(resp.h_q), $s_{j_q h_q}$ définira la contribution brute de la q-ème variable à la ressemblance entre o_i et $o_{i'}$.

Nous commencerons par déterminer la contribution réduite d'une même variable c_q (dont l'ensemble des modalités est codé par J_q) à la similarité entre deux objets, puis -de façon égale et parallèle- nous intégrerons l'ensemble des variables.

VI.2. Contribution de J_q à la ressemblance entre deux objets

L'indice q restant fixé dans ce paragraphe, nous l'omettrons pour des raisons de simplicité d'écriture.

L'élaboration de l'indice obéit dans notre approche à un principe statistique général de construction, où à partir d'un premier indice, localement défini, une normalisation est effectuée à partir de la distribution empirique de cet indice sur l'ensemble $P_2(E)$ des paires d'objets (ou parties à deux éléments) de E.

x et y désignant les deux objets à comparer, si $c(x)=j_o$ et $c(y)=h_o$, l'indice sera localement défini par le nombre $s_{j_o h_o}$ de la table (4). Il y a lieu par conséquent de préciser la distribution de $\{s_{jh} / (j,h) \in H\}$ sur $P_2(E)$. Cette distribution s'obtient très aisément à partir de la décomposition de $P_2(E)$ conformément à la partition de E déterminée par la variable qualitative c.

Plus directement, en désignant par n(j) le cardinal de la classe d'objets E_j, possédant la j-ème modalité du caractère c dont nous désignons par m le nombre de modalités,

$$\text{card}(P_2(E)) = \sum_{1 \leq j \leq m} n(j)(n(j)-1)/2 + \sum_{1 \leq j < h \leq m} n(j)n(h). \quad (5)$$

$$= n(n-1)/2.$$

Désignons respectivement par

$$\rho_j = n(j)(n(j)-1)/n(n-1) \text{ et } \sigma_{jh} = 2n(j)n(h)/n(n-1), \quad (6)$$

la proportion de paires d'objets $\{x',y'\}$ dont les deux composantes sont dans la classe E_j et celle pour lesquelles x' est dans la classe E_j et y' dans celle E_h, $1 \leq j \leq m$ et $1 \leq j < h \leq m$.

Désignons encore par

$$\rho = \sum_{1 \leq j \leq m} \rho_j \text{ et } \sigma = \sum_{1 \leq j < h \leq m} \sigma_{jh}, \quad (7)$$

qui sont respectivement, la proportion de paires réunies et séparées par la partition $\{E_j / 1 \leq j \leq m\}$.

Toutes les paires d'objets appartenant à E_j (resp. $E_j * E_h$) ont la même valeur s_{jj} (resp. s_{jh}) de l'indice local.

La distribution de la table des similarités $\{s_{jh} / (j,h) \in H\}$ sur $P_2(E)$ est donc définie par

$$\{(s_{kk},\rho_k),(s_{jh},\sigma_{jh}) / k \in J \text{ et } (j,h) \in J^{\{2\}}, \quad (8)$$

où nous avons noté $J^{\{2\}} = \{(j,h) / 1 \leq j < h \leq m\}$.

Calcul de la moyenne et de la variance de la distribution (8)

Si μ et μ_2 désignent la moyenne et le moment absolu d'ordre 2, on a :

$$\mu = \sum_{1 \leq k \leq m} \rho_k s_{kk} + \sum_{1 \leq j < h \leq m} \sigma_{jh} s_{jh}, \quad (9)$$

$$\mu_2 = \sum_{1 \leq k \leq m} \rho_k s_{kk}^2 + \sum_{1 \leq j < h \leq m} \sigma_{jh} s_{jh}^2 \quad (10)$$

On suppose -ce qui est naturel- que s_{kk} est le même pour tout $k=1,2,\ldots,m$. Notons s cette valeur commune.

$$\mu = \rho s + \sum_{1 \leq j < h \leq m} \sigma_{jh} s_{jh}, \quad (9')$$

$$W_2 = \rho s^2 + \sum_{1<j<h<m} \sigma_{jh} s_{jh}^2 \quad (10')$$

Le calcul informatique de la variance que nous notons ici w, utilise

$$w = W_2 - \mu^2 \quad (11)$$

Plus directement,

$$w = \rho(\sigma s - \sum_{j<h} \sigma_{jh} s_{jh})^2$$

$$+ \sum_{1<j<h<m} \sigma_{jh} (s_{jh} - \rho s - \sum_{\ell<k} \sigma_k s_{\ell k})^2, \quad (12)$$

qui se met sous la forme

$$= \sum_{1<j<h<m} \sigma_{jh} \left[\sum_{1<\ell<k<m} \sigma_{\ell k} (s_{jh} - s_{\ell k}) \right]^2, \quad (13)$$

en ayant au préalable noté σ_{kk} pour ρ_k, $1<k<m$.

Le numérateur de l'indice d'association s'écrit :

$$s_{j_o h_o} - \rho s - \sum_{1<j<h<m} \sigma_{jh} s_{jh}, \quad (14)$$

qu'on peut mettre sous la forme

$$\sum_{1<\ell<k<m} \sigma_{\ell k} (s_{j_o h_o} - s_{\ell k}). \quad (15)$$

L'indice d'association entre les deux objets x et y, relativement à la variable en question est égal à

$$S(x,y) = \frac{\sum_{1<\ell<k<m} \sigma_{\ell k} (s_{j_o h_o} - s_{\ell k})}{\left\{ \sum_{j<k} \sigma_{jh} \left[\sum_{\ell<k} \sigma_{\ell k} (s_{jh} - s_{\ell k}) \right]^2 \right\}^{1/2}}. \quad (16)$$

Nous allons à présent établir une propriété d'invariance de cet indice lorsque le nombre de modalités de la variable qualitative est deux.

Dans ce cas là, on a

$H=\{(1,1),(1,2),(2,2)\}$ et on posera

$s_{12}=p$, $s_{11}=s_{22}=q$, où $p<q$.

Dans ces conditions, la moyenne et la variance de la distribution des indices locaux sont respectivement égaux à

$$\mu = p\sigma + q\rho$$
$$\mathcal{W} = p^2\sigma + q^2\rho - (p\sigma - q\rho)^2 \qquad (17)$$

Si les deux objets à comparer possèdent deux modalités distinctes, la valeur de l'indice S s'écrit :

$$\frac{p-(p\sigma+q\rho)}{\sqrt{p^2\sigma+q^2\rho-(p\sigma+q\rho)^2}} \qquad (18)$$

qui -après calcul- se réduit à

$$-\sqrt{\rho/\sigma} \qquad (19)$$

Si les deux objets à comparer possèdent la même modalité, on remplacera le numérateur de (18) par $\left[q-(p\sigma+q\rho)\right]$ et l'indice S se réduit à

$$+\sqrt{\sigma/\rho}\ . \qquad (20)$$

D'où l'énoncé du résultat :

Propriété. Si le nombre de modalités de la variable qualitative se réduit à deux, l'indice globalement réduit ne dépend plus que de la répartition des deux modalités sur l'ensemble des objets.

De façon précise, si n(1)(resp.n(2)) est le nombre d'objets possédant la modalité 1 (resp.2), deux objets x et y possédant respectivement les modalités 1 et 2, ont pour valeur de l'indice S :

$$S(x,y) = \sqrt{\frac{1}{2}\ \frac{(n(1)-1)}{n(2)} + \frac{(n(2)-1)}{n(1)}}\ . \qquad (19')$$

Si par contre les deux objets ont la même modalité :

$$S(x,y) = +\sqrt{2\ \left\{\frac{n(1)n(2)}{n(1)(n(1)-1)n(2)(n(2)-1)}\right\}}. \qquad (20')$$

Cette propriété qui peut surprendre est en fait heureuse et naturelle : la perception des ressemblances mutuelles entre objets pris dans un même ensemble, face à une simple dichotomie, n'a plus à dépendre d'une "quantification" de cette dichotomie.

Considérons deux objets face à plusieurs variables dichotomiques. Les formules (19') et (20') montrent que la contribution d'une même variable à la ressemblance des deux objets qui en possèdent la même modalité (resp. qui n'en possèdent pas la même modalité) est d'autant plus élevée que les deux modalités se trouvent plus également réparties.

La propriété d'invariance ci-dessus -par rapport à la table (4) des similarités locales- n'est plus valable si la variable a plus de deux modalités.

Pour la plupart des applications, une information très générale de type "préordonnance" sur H est suffisamment fine pour une excellente reconnaissance des classes de proximité sur l'ensemble des objets.

La prise en compte de variables "préordonnance" permet d'enrichir sensiblement la structure descriptive des variables qualitatives et ce, à partir de la perception a priori de la ressemblance entre modalités d'une même variable. Ainsi, même dans le cas le plus pauvre d'une variable logique où a(resp.$\bar{a}$) désigne la présence (resp. absence) de l'attribut, on peut définir la préordonnance : $a\bar{a} \langle \bar{a}\bar{a} \langle aa$, où la présence commune est plus indicative de la ressemblance que l'absence commune.

VI.3. Indice d'association dans le cas de plusieurs variables

Désignons par $S_q(x,y)$ la contribution de la variable c_q à la comparaison des deux objets x et y. $S_q(x,y)$ est donné par la formule (16) ci-dessus relativement à la table (4).

Pour un même q, la moyenne et la variance de S_q sur l'ensemble $P_2(E)$ sont respectivement égales à 0 et à 1.

Pour la définition de l'indice d'association, on tiendra également compte des différentes variables en proposant

$$S(x,y) = \frac{1}{\sqrt{Q}} \sum_{1 \langle q \langle Q} S_q(x,y). \quad (21)$$

Rappelons que la réduction au moyen de $1/\sqrt{Q}$ se réfère à un modèle d'indépendance où les v.a. associés aux c_q, $1 \langle q \langle Q$, ont une variance unité.

Encore une fois, on reprendra ici le dernier alinéa du paragraphe III.

VI.4. Une solution efficace et simple au problème du consensus entre arbres de classifications

Ces dernières années, on a vu se développer un intérêt tout particulier pour le problème de la recherche d'un consensus entre arbres de classifications sur le même ensemble (ROHLF(1982)), (DIDAY(1982)), (BARTHELEMY, LECLERC, MONJARDET(1984)),... Etant donné un ensemble fini d'arbres de classifications ou hiérarchies indicées $H_1, H_2, \ldots, H_Q$, sur le même ensemble fini E d'objets, il s'agit de résumer "au mieux" cet ensemble de hiérarchies au moyen d'un arbre unique de classifications par rapport auquel on pourra situer les différents arbres.

Le problème concret peut se présenter dans le cas d'un tableau de données ExV (E:ensemble des objets et V:ensemble des variables descriptives), où les variables sont mesurées à différentes dates. Pour chaque date, le tableau des mesures (indexé par ExV) conduit -par une méthode de classification fixée- à un arbre de classification sur E. Pour une période relativement stable, on peut vouloir -sur la seule base des arbres obtenus- une organisation classificatoire hiérarchique "moyenne" de l'ensemble E des objets.

Une autre situation concrète est celle où on dispose de Q ensembles de variables $V_1, V_2, \ldots, V_Q$, mesurées sur le même ensemble E des objets et où, relativement à un même tableau de données ExV_q, on suppose construit un arbre de classification H_q sur E. On se pose alors le problème de résumer $\{H_q / 1 \leq q \leq Q\}$ au moyen d'un seul arbre H de classification.

Nous allons commencer par évoquer le problème dans le cas particulier où la donnée est une suite de partitions $\{P_q / 1 \leq q \leq Q\}$ sur E. En effet, la donnée d'une partition P est équivalente à celle d'un arbre à trois niveaux définissant la suite des partitions P_0, P et P_1, où P_0 et P_1 sont respectivement la partition discrète à n classes et celle grossière à une seule classe.

Un des initiateurs du problème de la recherche d'un consensus d'une suite $\{P_q / 1 \leq q \leq Q\}$ de partitions au moyen d'une seule partition qu'il appelle "centrale" est S. Régnier (REGNIER(1965)). Mais le problème qui nous occupe ici est de fournir ce consensus sous la forme sensiblement plus riche d'un arbre hiérarchique des classifications que -dans notre méthode- on condense aux niveaux où apparaît un noeud significatif (LERMAN(1970a),(1973),(1981), (1983a)).

Nous résolvons bien ce problème au niveau du paragraphe IV ci-dessus, ainsi d'ailleurs qu'au niveau du paragraphe VI.2 qui précède, mais dans le cas où les variables qualitatives sont chacune à deux modalités et avec une représentation en termes de préordonnance.

Cette dernière représentation va nous permettre de proposer une nouvelle solution -par rapport à celle du paragraphe IV ci-dessus- pour le problème du consensus, au moyen d'un arbre hiérarchique des classifications (i.e. chaîne ordonnée de partitions sur E), d'une famille de partitions.

Cette solution représentera un cas particulier de celui général où la donnée est une famille $\{H_q/1\leq q\leq Q\}$ de chaînes de partitions ou arbres de classifications.

VI.4.1. Cas où la donnée est une famille de partitions

Désignons par P l'une des partitions de la famille $\{P_q/1\leq q\leq Q\}$ et par F l'ensemble $P_2(E)$ des parties à deux éléments de l'ensemble des objets. Nous représentons P au niveau de F au moyen d'un préordre à deux classes S et R où S(resp. R) désigne l'ensemble des paires séparées (resp. réunies) par la partition P. Si on désigne par $(n_1,n_2,\ldots,n_j,\ldots,n_k)$ la suite des cardinaux des classes $\{E_j/1\leq j\leq k\}$, on a, avec des notations que l'on comprend,

$$R=\sum_{1\leq j\leq k} P_2(E_j) \text{ et } r=\mathrm{card}(R)=\sum_{1\leq j\leq k} n_j(n_j-1)/2, \quad (22)$$

où la première somme est ensembliste. D'autre part,

$$S=\sum_{1\leq j<h\leq k} E_j * E_h \text{ et } s=\mathrm{card}(S)=\sum_{1\leq j<h\leq k} n_j n_h, \quad (23)$$

où -rappelons-le- $E_j * E_h$ désigne l'ensemble des paires d'objets dont l'un des éléments appartient à E_j et l'autre à E_h. Enfin,

$$f=\mathrm{card}(F)=r+s. \quad (24)$$

En supposant bien entendu S≺R pour l'ordre quotient, le rang d'une même paire p -conformément à la fonction ordinale rappelée au paragraphe IV.1 est

$(s+1)/2$ si $p\in S$ et $s+(r+1)/2$ si $p\in R$, soit

$(f-r+1)/2$ si $p\in S$ et $(2f-r+1)/2$ si $p\in R$.

Nous allons maintenant déterminer la contribution normalisée de la partition P à la comparaison de deux objets donnés x et y.

Moyenne et Variance du rang d'une paire

La moyenne peut s'écrire

$$\frac{1}{f}\left((f-r)(f-r+1)/2+r(2f-r+1)/2\right)$$

$$= \frac{1}{2}(f+1). \quad (25)$$

La variance se met sous la forme

$$\frac{1}{f}\left((f-r)r^2/4+r(f-r)^2/4\right)=r(f-r)/4. \quad (26)$$

Dans ces conditions, la contribution à la ressemblance entre x et y de la partition P, est égale à

$$\left((f-r+1)/2-(f+1)/2\right)/\sqrt{r(f-r)/4}=-\sqrt{r/s}, \quad (27)$$

si $p=\{x,y\}$ appartient à S et

$$\left((2f-r+1)/2-(f+1)/2\right)/\sqrt{r(f-r)/4}=+\sqrt{s/r}, \quad (28)$$

si $p=\{x,y\}$ appartient à R.

On obtient ainsi exactement le même résultat que dans le cas où la partition a exactement deux classes (cf. formules (19) et (20) du paragraphe IV.2. ci-dessus).

Si on désigne par r_q (resp.s_q) le nombre de paires réunies (resp. séparées) par la partition P_q, $1 \leqslant q \leqslant Q$, l'indice de similarité S(x,y) entre les deux objets x et y tenant compte de la famille de partitions -définies en l'occurrence par des variables qualitatives nominales- se met sous la forme :

$$S(x,y)= \frac{1}{\sqrt{Q}} \sum_{1<q<Q} \phi_q(x,y), \quad (29)$$

où $\phi_q(x,y)=-\sqrt{r_q/s_q}$ (resp.$+\sqrt{s_q/r_q}$) si $\{x,y\}$ appartient à S_q (resp.R_q) où on comprend que S_q (resp. R_q) est l'ensemble des paires séparées (resp. réunies) par la partition P_q, $1<q<Q$.

La table des indices (29) :

$$\{S(x,y)/\{x,y\} \in P_2(E)\}, \quad (30)$$

correspondra à celle (14) du paragraphe II, pour subir les mêmes transformations et être assimilée par l'algorithme de la vraisemblance du lien.

VI.4.2. Cas où la donnée est une famille de chaînes de partitions

La donnée est ici une famille $\{H_q/1\leq q\leq Q\}$ d'arbres des classifications ou hiérarchies indicées sur l'ensemble E des objets. Nous désignons par H l'une quelconque d'entre elles dont nous allons préciser la contribution normalisée à la ressemblance de deux objets donnés x et y de E.

La détermination repose sur une équivalence que nous avons établie dans (LERMAN(1970a)) entre la donnée d'un arbre de classifications sur E et celle d'une préordonnance sur E d'un type particulier que nous appelons "ultra-métrique" et dont nous allons -tant soit peu- rappeler la construction.

Une préordonnance $\mathcal{O}(E)$ sur E est un préordre total sur $P_2(E)$ qui est sensé refléter de façon ordinale les ressemblances mutuelles entre éléments de E. On suppose que ce préordre est établi de telle sorte qu'une même paire $p=\{x,y\}$ se situe d'autant plus à droite (i.e. à un rang d'autant plus élevé) que la ressemblance entre les composantes x et y est plus grande.

La préordonnance $\mathcal{O}(E)$ est ultramétrique si et seulement si, quel que soit l'élément $\{x,y,z\}$ de l'ensemble $P_3(E)$ des parties à trois éléments de E, chacune des trois paires $\{x,y\}$, $\{x,z\}$ et $\{x,z\}$ est à droite de celle des deux autres paires la plus à gauche. Plus précisément, si on appelle ρ une fonction ordinale compatible avec le préordre, qu'on peut directement prendre définie comme il est exprimé au paragraphe VI.1., la condition s'écrit

$$(\forall\{x,y,z\} \in P_3(E)),\ \rho(x,y)\geq\min(\rho(x,z),\rho(y,z)) \ . \quad (31)$$

Relativement à un arbre de classifications qui définit une chaîne ordonnée de partitions $(P_0,P_1,\ldots,P_i,\ldots,P_I)$ où P_0 est la partition discrète (à n classes) et P_I celle grossière (à une classe), la préordonnance ultramétrique associée $\mathcal{O}(E)$ peut se construire pas à pas de la manière suivante :

Si on note $\{E_j^{(I-1)}/1\leqslant j\leqslant k(I-1)\}$ la partition $P_{(I-1)}$, le passage de $P_{(I-1)}$ à P_I résulte de la fusion en une seule classe E, des classes $E_j^{(I-1)}$, $1\leqslant j\leqslant k(I-1)$. Dans ces conditions, l'ensemble des paires de la première classe de la préordonnance est défini par

$$\sum_{1\leqslant j\leqslant h\leqslant k(I-1)} E_j^{(I-1)} * E_h^{(I-1)} \quad ; \quad (32)$$

en d'autres termes, une paire $p=\{x,y\}$ appartient à la première classe de la préordonnance ultramétrique si et seulement si ses deux composantes appartiennent respectivement à deux classes distinctes de $P_{(I-1)}$.

Le passage de $P_{(I-2)}$ à $P_{(I-1)}$ résulte de la réunion de certaines classes de $P_{(I-2)}$. La deuxième classe de la préordonnance est formée de l'ensemble des paires d'objets dont les deux composantes appartiennent respectivement à deux classes distinctes de $P_{(I-2)}$ qui ont fusionné dans le passage de $P_{(I-2)}$ à $P_{(I-1)}$.

Et ainsi de suite, de la même façon, le passage de $P_{(I-3)}$ à $P_{(I-2)}$ détermine la troisième classe de la préordonnance ultramétrique,...,celui de P_0 à P_1, la I-ème classe de la préordonnance.

Revenons à présent au début de ce sous-paragraphe et associons à chaque H_q la préordonnance ultramétrique ϖ_q sur E comme nous venons tout juste de l'exprimer. La donnée équivalente est donc maintenant $\{\varpi_q/1\leqslant q\leqslant Q\}$ dont nous désignons par ϖ un élément quelconque.

La contribution normalisée de ϖ à la ressemblance de deux objets x et y se fait de façon analogue à celle du paragraphe VI.2., mais plus spécifique.

Plus précisément, ρ désignant la fonction rang (cf. § VI.1.) associée à ϖ, on rapporte $\rho(\{x,y\})$ à la moyenne M et à la variance V de $\{\rho(p)/p\in F\}$. On a donc à calculer

$$\sum\{\rho(p)/p\in F\} \text{ et } \sum\{\rho^2(p)/p\in F\} \qquad (33)$$

qui dépendent directement du type de la hiérarchie H associée à ϖ; c'est-à-dire, des cardinaux des classes des différentes partitions.

Comme d'habitude, cette contribution normalisée s'écrit :

$$\frac{\rho(\{x,y\})-M}{\sqrt{V}} . \qquad (34)$$

La prise en compte de la famille $\{\varpi_q / 1 \leq q \leq Q\}$ conduit à l'indice

$$S(x,y) = \frac{1}{\sqrt{Q}} \sum_{1 \leq q \leq Q} \frac{\rho_q(\{x,y\})-M_q}{\sqrt{V_q}}, \qquad (35)$$

où ρ_q, M_q et V_q sont calculés par rapport à ϖ_q, de la même manière que ρ, M et V le sont par rapport à ϖ.

Il nous reste à ajouter l'équivalent du dernier alinéa du paragraphe VI.4.1. ci-dessus, pour offrir au moyen de l'algorithme de la vraisemblance du lien, un consensus raffiné des différents arbres de classifications $H_1, H_2, \ldots, H_q, \ldots, H_Q$.

VII. CAS OU LES VARIABLES SONT DE TYPES DIVERS

Dans les précédents paragraphes les variables de description de l'ensemble des objets sont toutes d'un seul type et nous avons pris en considération six cas différents : le "numérique" (n), le "logique" (ℓ), le "qualitatif nominal" (qn), le "qualitatif ordinal" (qo) et le "qualitatif préordonnance" (pr).

Nous supposons ici que dans le cadre d'un même tableau de données Objets x Variables, on rencontre des variables de types différents. Ainsi, l'ensemble V des variables est supposé pouvoir être décomposé en la somme ensembliste :

$$V = V_n + V_\ell + V_{qn} + V_{qo} + V_{pr}, \qquad (1)$$

où, respectivement, V_n, V_ℓ, V_{qn}, V_{qo} et V_{pr} sont les ensembles de variables numériques, logiques (i.e. attributs), qualitatives nominales, qualitatives ordinales et préordonnances.

L'intérêt de notre indice est de procéder de façon additive, variable par variable, par contributions normalisées. Toutefois, dans les cas numérique et logique où les variables sont -en se plaçant d'un point de vue "géométrique"- de caractère unidimensionnel et "orienté" il y a lieu au préalable de connecter la mesure d'une variable sur un objet à celle des autres variables de même type.

D'où, la construction suivante de l'indice global pour la comparaison de deux objets o_i et $o_{i'}$:

$$S(o_i,o_{i'}) = \frac{1}{\sqrt{card(V)}} \{\sum (s_{V_n}(o_i,o_{i'}) - M^{V_n})/\sigma^{V_n} | v_n \in V_n\}$$

$$+ \sum\{(s_{\alpha v_\ell}(o_i,o_{i'}) - M_\alpha^{v_\ell})/\sigma_\alpha^{v_\ell} | v_\ell \in V_\ell\}$$

$$+ \sum\{(s_{V_{qn}}(o_i,o_{i'}) - M^{V_{qn}})/\sigma^{V_{qn}} | v_{qn} \in V_{qn}\}$$

$$+ \sum\{(s_{V_{qo}}(o_i,o_{i'}) - M^{V_{qo}})/\sigma^{V_{qo}} | v_{qo} \in V_{qo}\}$$

$$+ \sum\{(s_{V_{pr}}(o_i,o_{i'}) - M^{V_{pr}})/\sigma^{V_{pr}} | v_{pr} \in V_{pr}\} , \quad (2)$$

où les notations sont assez claires pour avoir été maintes fois utilisées. Plus précisément, l'indice comprend cinq sommes ; la première correspond à celle (13) du paragraphe II où V se trouve remplacé par V_n, la seconde correspond à celle (7) du paragraphe III où V est remplacé par V_ℓ, la troisième correspond à celle (6) du paragraphe IV où V est remplacé par V_{qn}, la quatrième correspond à celle (8) du paragraphe V où V est remplacé par V_{qo}, enfin, la cinquième correspond à celle (21) du paragraphe VI où l'ensemble des variables est défini par V_{pr}.

Reprenons ici le problème de la construction d'un arbre de classification sur un ensemble d'objets décrits par plusieurs ensembles de variables. A ce niveau et par rapport au paragraphe VI.4.2. précédent, la solution se présente sous la forme d'une alternative ayant un caractère fondamental.

Le premier terme, le plus simple, correspond à intégrer les différents ensembles de variables (de même type ou de types différents) dans le cadre d'un même indice tel que (2) pour la comparaison deux à deux des objets.

Le deuxième terme de l'alternative correspond à associer à chaque ensemble de variables, un arbre de classification sur l'ensemble des objets, puis à réaliser le consensus des différents arbres obtenus, conformément au traitement du paragraphe VI.4.2.

VIII. CAS D'UN SEUL TABLEAU PUIS D'UNE JUXTAPOSITION DE TABLEAUX DE CONTINGENCE

VIII.1. Cas d'un seul tableau

Un tableau de contingence est un tableau de croisement entre deux variables qualitatives nominales. En désignant par I (resp.J) l'ensemble des modalités de la première (resp. seconde), ce tableau se met sous la forme :

$$\{k_{ij}/(i,j) \in I \times J\}, \quad (1)$$

où k_{ij} désigne le nombre d'individus possédant les modalités i de I et j de J. On rappelle les notations :

$$k_{i.}=\sum\{k_{ij}/j\in J\},\ k_{.j}=\sum\{k_{ij}/i\in I\}$$

et

$$k_{..}=\sum\{k_{i.}/i\in I\}=\sum\{k_{.j}/j\in J\}. \quad (2)$$

Relativement au problème de l'analyse de I à travers J (resp.J à travers I), cette structure parfaitement symétrique du tableau des données occupe une position pivot. On sait très bien qu'elle a nourri l'analyse factorielle des correspondances qui lui a trouvé la représentation euclidienne adéquate.

Pour fixer les idées, considérons le problème de la classification hiérarchique de I à travers J et introduisons les deux nuages classiques de R^m et de R^n (m=card(J) et n=card(I))

$$N(I)=\{(f_J^i, p_{i.})/i\in I\} \text{ et } N(J)=\{(f_I^j, p_{.j})/j\in J\}, \quad (3)$$

où on rappelle que le point f_J^i de $R^{|J|}$ (resp. f_I^j de $R^{|I|}$) qui définit le profil de i à travers J (resp. de j à travers I) a comme suite de coordonnées $(k_{ij}/k_{i.}\,|\,1\leq j\leq m)$ [resp. $(k_{ij}/k_{.j}\,|\,1\leq i\leq n)$]. On rappelle également que

$\{1/p_{.j}$ où $p_{.j}=k_{.j}/k_{..} \mid 1\leq j\leq m\}$ (resp.$\{1/p_{i.}$ où $p_{i.}=k_{i.}/k_{..} \mid 1\leq i\leq n\}$) déterminent les métriques diagonales du χ^2 dont on munit R^m et R^n pour l'analyse de N(I) (resp.N(J)).

Pour le problème posé, la classification des moindres carrés -conformément à la métrique du χ^2- utilise directement le nuage N(I) de l'espace euclidien R^m où -soulignons-le- chaque i de I se trouve interprété comme un point-objet (BENZECRI(1973)).

Jusqu'à présent, dans notre approche basée sur la vraisemblance du lien, nous avons au contraire été conduits -pour le problème posé de la classification de I- à utiliser le nuage N(J) en interprétant chaque i de I comme une variable dont la mesure sur le point-objet j est f_i^j. Dans ces conditions, l'indice d'association entre i et i' de I a la nature d'une corrélation dont -avec B. TALLUR- nous avons donné une expression et une interprétation géométrique (LERMAN & TALLUR(1980)), puis une expression et interprétation ensembliste et statistique (LERMAN(1983b)).

Ce que nous proposons ici -toujours dans le cadre de notre approche- c'est de reprendre N(I) et d'établir la table des indices d'association

$$\{S(i,i')/\{i,i'\} \in P_2(I)\} \qquad (3)$$

sur la même base que les indices développés dans ce texte et en accord avec la représentation euclidienne que nous venons de rappeler.

Toutefois, on se rendra compte avec intérêt qu'un indice type "cosinus" dans l'un des espaces s'interprète comme un indice type "corrélation" dans l'autre espace.

1- Contribution de j à Cos(i,i')

Nous considérons comme point de référence le cas numérique (cf. § II) et le coefficient d'association dont le point de départ est la contribution de la j-ème composante à l'indice cosinus. On a

$$\langle f_J^i-0, f_J^{i'}-0\rangle = \sum_{1\leq j\leq m} \frac{f_j^i f_j^{i'}}{p_{.j}}, \qquad (4)$$

où $\langle .,. \rangle$ désigne le produit scalaire. D'autre part,

$$||f_J^i - 0||^2 = \sum_{1 \leq j \leq m} \frac{(f_j^i)^2}{p_{.j}} . \qquad (5)$$

Dans ces conditions, on obtient

$$\text{Cos}(i,i') = \frac{\sum_j (f_{ij} f_{i'j} / p_{.j})}{\sqrt{(\sum_j (f_{ij}^2 / p_{.j}))(\sum_h (f_{i'h}^2 / p_{.h}))}} . \qquad (6)$$

Dans ces conditions, la contribution de j à Cos(i,i') est :

$$s_j(i,i') = \frac{(f_{ij} f_{i'j} / p_{.j})}{\sqrt{(\sum_h (f_{ih}^2 / p_{.h}))(\sum_k (f_{i'k}^2 / p_{.k}))}} . \qquad (7)$$

2- Moyenne et Variance sur $P_2(I)$ de $s_j(i,i')$. Proposition de l'indice de similarité

La moyenne et la variance tiennent compte de la distribution $\{p_{i.} / i \in I\}$ de poids dont se trouve muni I.

Dans ce cas, on a pour la moyenne :

$$M^j = \frac{2}{(1 - \sum_i p_{i.}^2)} \sum \{p_{i.} p_{i'.} \, s_j(i,i') / \{i,i'\} \in P_2(I)\}. \qquad (8)$$

De même, le moment absolu d'ordre 2 prend la forme :

$$M_2^j = \frac{2}{(1 - \sum_i p_{i.}^2)} \sum \{p_{i.} p_{i'.} (s_j(i,i'))^2 / \{i,i'\} \in P_2(I)\}. \qquad (9)$$

D'où la variance

$$\sigma_j^2 = (M_2^j - M_j^2) \qquad (10)$$

et la contribution normalisée de j à la comparaison de i et de i' :

$$S_j(i,i') = \frac{(s_j(i,i') - M^j)}{\sigma^j}. \quad (11)$$

L'indice d'association global -tenant compte de l'ensemble J- prend comme d'habitude une forme additive :

$$S(i,i') = \sum_{1 \leq j \leq m} S_j(i,i'), \quad (12)$$

ce qui précise l'élément courant de la table (3) ci-dessus, laquelle subira la même suite de transformations que celle (14) du paragraphe II, avant application de l'algorithme de la vraisemblance du lien.

3- Identité entre un indice type cosinus dans l'un des espaces et un indice type corrélation dans l'autre espace.

Si v et w sont deux variables numériques ordinaires, nous désignerons par

$$Cor_0(v,w) = \frac{m_{11}(v,w)}{\sqrt{m_2(v)m_2(w)}}, \quad (13)$$

où $m_2(v)$ (resp.$m_2(w)$) est le moment absolu (par rapport à l'origine) de la variable v(resp.w) et où $m_{11}(v,w)$ est le moment produit absolu entre v et w.

Cor(v,w) indique bien entendu le coefficient de corrélation ordinaire qui correspond à la même expression (13) mais où les moments sont centrés (par rapport aux moyennes des variables v et w).

Relativement à la représentation du nuage N(J) où I est assimilé à l'ensemble des variables et où la mesure de i ($\in$I) sur j ($\in$J) est f_i^j (LERMAN-TALLUR(1980)), on a après simplification :

$$Cor_0(i,i') = \frac{\sum_j (f_{ij} f_{i'j} / p_{.j})}{\sqrt{\left(\sum_j (f_{ij}^2 / p_{.j})\right)\left(\sum_h (f_{i'h}^2 / p_{.h})\right)}}, \quad (14)$$

qui n'est autre que Cos(i,i') calculé dans le cadre de la représentation du nuage N(I). D'où l'identité

Propriété 1. $Cor_0(i,i')/N(J) = Cos_0(i,i')/N(I)$. (15)

Nous allons maintenant examiner, relativement à la représentation de N(I), le cosinus de l'angle $\widehat{f_J^i\ g_J\ f_J^{i'}}$, où g_J est le centre de gravité du nuage.

En d'autres termes, il s'agit de

$$\langle f_J^i - g_J, f_J^{i'} - g_J \rangle / \| f_J^i - g_J \| \cdot \| f_J^{i'} - g_J \| , \quad (16)$$

que nous pouvons noter $Cos_g(i,i')$. Après calcul et simplification, on obtient :

$$Cos_g(i,i') = \frac{(\sum_j (f_{ij}f_{i'j}/p_{.j}) - p_{i.}p_{i'.})}{\sqrt{(\sum_j (f_{ij}^2/p_{.j}) - p_{i.}^2)(\sum_h (f_{i'h}^2/p_{.h}) - p_{i'.}^2)}} , \quad (17)$$

qui n'est autre que le coefficient de corrélation entre i et i', relativement à la représentation de N(J). D'où

Propriété 2. $Cor(i,i')/N(J) = Cos_g(i,i')/N(I)$. (18)

Il en résulte -conformément au point de vue développé dans cette étude- la possibilité de proposer un autre indice, à partir des contributions élémentaires des différents j de J. Plus précisément, dans les calculs précédents (cf. § 2 ci-dessus), on substituera à $s_j(i,i')$, l'expression

$$s_j'(i,i') = \frac{(f_j^i - p_{.j})(f_j^{i'} - p_{.j})/p_{.j}}{\sqrt{(\sum_h (f_h^i - p_{.h})^2/p_{.h})(\sum_k (f_k^{i'} - p_{.k})^2/p_{.k})}} , \quad (19)$$

qui représente la contribution de j à l'expression (16).

Si nous n'avions pas dans le cas numérique (cf.§II) considéré l'indice "cosinus" à partir du centre de gravité du nuage, c'est que d'une part, les justifications formelle et statistique n'étaient pas claires et que surtout, on obtenait de meilleurs résultats avec l'indice (2)(§ II). Alors que pour la structure des données qui nous concerne ici, l'indice (17) a donné d'excellents résultats.

VIII.2. Cas d'une juxtaposition "horizontale" de tableaux

L'ensemble d'indexation d'une telle juxtaposition "horizontale" de tables de contingence, se met sous la forme

$$Ix(J^{(1)} \cup J^{(2)} \cup \ldots \cup J^{(\ell)} \cup \ldots \cup J^{(L)}), \qquad (20)$$

où I(resp. $J^{(\ell)}$, $1 \leq \ell \leq L$) se trouve défini par l'ensemble des modalités d'une variable-partition (i.e. qualitative nominale).

Compte tenu de la forme additive de l'indice, nous avons ici peu de choses à ajouter. Mais, si nous avons considéré cette forme du tableau des données, c'est qu'elle se rencontre fréquemment dans les données géographico-économiques.

De façon précise, pour la comparaison de i et i' de I, à chaque table de contingence $IxJ^{(\ell)}$, nous associons l'indice (12) où celui qu'on peut déduire de (19),

$$S_\ell(i,i') = \sum_{1 \leq j_\ell \leq m} S_{j_\ell}(i,i'), \qquad (21)$$

où on a noté $m_\ell = \mathrm{card}(J^{(\ell)})$.

Pour tenir compte de l'ensemble des $J^{(\ell)}$, $1 \leq \ell \leq L$, en neutralisant l'influence du nombre de modalités de la variable qualitative pour la comparaison de i et de i', on peut adopter l'indice

$$S = \sum_{1 \leq \ell \leq L} \frac{1}{m_\ell} S_\ell(i,i'). \qquad (13)$$

IX. CONCLUSION

La plupart des indices exprimés ici ont été testés et en passe d'être développés dans le cadre d'un élégant programme (cf. partie suivante du rapport). Ces indices ont montré toute leur pertinence par rapport à ceux que nous utilisions précédemment et qui nous ont d'ailleurs conduit à notre actuelle démarche.

Ainsi, avec la présente étude, nous enlevons une grande indétermination quant au choix de l'indice de ressemblance entre objets, compatible avec l'algorithme de la vraisemblance du lien. De la sorte, la méthode de classification hiérarchique basée sur la vraisemblance des liens atteint un très grand niveau d'achèvement, puisqu'elle permet, quel que soit la structure du tableau des données et avec une très grande fidélité dans la représentation mathématique, de classifier l'ensemble des variables (LERMAN(1981)) ainsi que l'ensemble des objets. Une vue générale de l'étude du cas spécifique mais important, de la juxtaposition de tables de contingence, est fournie dans (LERMAN(1984)). Un complément intéressant à cette étude est fourni au paragraphe précédent avec le point de vue qui prédomine dans ce texte.

BIBLIOGRAPHIE

BARTHELEMY J.P., LECLERC B. et MONJARDET B. (1984) ; "Quelques aspects du consensus en classification" in "Data Analysis and Informatics", North Holland.

BENZECRI J.P. & Collaborateurs (1973) ; "L'analyse des Données, Tome I : La Taxinomie, IB n°5", Dunod, Paris.

CHAH S. (1984) ; "Agrégation des préordonnances", Etude F-063, Centre scientifique IBM de Paris.

DIDAY E. (1982) ; "Croisements, ordres et ultramétriques : applications à la recherche de consensus en classification automatique", Rap. de rech. n°144, I.N.R.I.A.

GOWER J.C. (1971) ; "A general coefficient of similarity and some of its properties", Biometrics, 27, pp. 857-872.

LERMAN I.C. (1970a) ; "Les bases de la classification automatique", Gauthier-Villars, "Collection Programmation", Paris.

LERMAN I.C. (1970b) ; "Sur l'analyse des données préalable à une classification automatique. Proposition d'un nouvel indice de similarité", Rev. Math et Sc. Hum. n°32, également paru dans "Mathematics in the Archaeological and Historical Sciences", Eddinburgh University Press, (1971).

LERMAN I.C. (1973) ; "Etude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la classification automatique", Cahiers du B.U.R.O. n°19 Paris.

LERMAN I.C. (1981) ; "Classification et analyse ordinale des données", Dunod, Paris.

LERMAN I.C. (1983a) ; "Sur la signification des classes issues d'une classification automatique", in "Numerical Taxonomy", Springer.

LERMAN I.C. (1983b) ; "Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence", Rev. Math. & Sc. Hum., 21è année, n°83, p. 5 à 30.

LERMAN I.C. (1984) ; "Analyse classificatoire d'une correspondance multiple, typologie et régression", in "Data Analysis and Informatics", North Holland.

LERMAN I.C. et TALLUR B. (1980) ; "Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence", Rev. de Stat. Appl. n°28,33, pp. 5-28, Paris.

MASSE J.R. (1978) ; "Classes de tableaux équivalents en analyse descriptive des données. Application à l'étude de mesures statiques sur circuits intégrés logiques", Thèse de 3ème cycle, Univ. de Rennes I.

OCHIAI A. (1957) ; "Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions", Bull. Jap. Soc. Sci. Fish., T 22, pp. 526-530.

REGNIER S. (1965) ; "Sur quelques aspects mathématiques des problèmes de classification automatique", I.C.C. Bulletin, Vol. 4 pp. 175-191.

ROHLF F.J. (1982) ; "Consensus indices for comparing classifications", Math. Biosc. 59, pp. 131-144,

SNEATH P.H.A. & SOKAL R. (1972) ; "Numerical Taxonomy", Freeman, San Francisco and London.

DOCUMENTATION TECHNIQUE

DU PROGRAMME S I M O B

0. NOM ET OBJET

Le programme SIMOB que nous présentons assure le calcul des indices de proximité entre individus décrits par différents types de variables. Les indices de proximité calculés sont ceux définis dans la partie théorique. Les variables descriptives traitées sont :

* Les variables numériques
* Les attributs
* Les variables partitions
* Les variables préordinales
* Les tableaux de contingence
* Les variables préordonnances

Version 0.0. : MAI 1985

1. PARAMETRES DE FONCTIONNEMENT

1.1. Cartes à fournir :

Dans le cas où un programme conversationnel ne se charge pas de remplir le fichier des paramètres, il faut introduire les cartes suivantes dans le fichier numéro 10.

Première carte : titre de l'expérience sur 80 caractères au maximum.
Deuxième carte : IPARAM(I), I=1,6 (format 6I5).
Troisième carte : Format des données sur 80 caractères au maximum.
Cartes suivantes : Uniquement pour les variables partitions, préordinales, préordonnances ou les tableaux de contingence : nombre de modalités par variables ou nombre de sous-tableaux suivant le cas (format 20I4).
Cartes suivantes : Uniquement pour les variables préordonnances : tableau des rangs NRG (format 20I4).

1.2. Description du vecteur des paramètres :

IPARAM(1) : nombre d'individus (NIND)
IPARAM(2) : nombre de variables (NVAR)
IPARAM(3) : type des variables (INDICE)

-) 1 : variables numériques
-) 2 : attributs
-) 3 : variables partitions
-) 4 : variables préordinales
-) 5 : tableau de contingence
-) 6 : variables préordonnances

IPARAM (4) : procédé de réduction (IRD)
dans le cas des attributs :
---) 1 : procédé 0 (voir partie théorique)
---) 2 : procédé 1
---) 3 : procédé 2
sinon :
---) 0

IPARAM (5) : dans le cas d'un tableau de contingence :
nombre de sous-tableaux ; (NST)
sinon : 0

IPARAM (6) : dans le cas de variables préordonnances :
NRGT : dimension du tableau des rangs
sinon : 0

$$NRGT = \sum_{i=1}^{nvar} (NBMØD(I)*(NBMØD(I)+1))/2 ;$$

où NVAR est le nombre de variables et NBMØD(I) est le nombre de modalités de la variable I.

1.3. Tableau des rangs NRG (variables préordonnances) :

Les rangs entre deux modalités d'une variable y sont rangés variable par variable et pour chaque variable l'adresse du rang entre les modalités i et j (avec j>i) est : debu-1+(j*(j-1))/2+i

où debu est l'adresse du premier rang dans le tableau NRG pour la variable.

2. ASPECTS INFORMATIQUES :

2.1. Gestion de mémoire :

Elle est effectuée par simulation d'allocation dynamique en piochant dans deux super-tableaux (INTG pour les entiers et REEL pour les réels). Le sous-programme responsable de cette gestion est ALLØC. Elle est complétée par le sous-programme ALLØ2 dans le cas de variables qualitatives (nominales ou ordinales), de variables préordonnances ou encore dans le cas d'un tableau de contingence.

Toutes les variables communes à plusieurs sous-programmes sont passées en paramètres (il n'y a ni espaces communs, ni équivalences).

2.2. Indications de performances :

Les mesures qui suivent ont été obtenues pour le calcul des indices de proximité sur un calculateur HB68 avec le système MULTICS.

	nombre d'individus	nombre de variables	temps
variables numériques	93	14	10.5 s
attributs	75	12	7.3 s
variables partitions	74	8	6.1 s
variables préordinales	74	8	6.5 s
variables préordonnances	93	14	17.3 s
tableau de contingence	89	14	20.1 s

3. SOUS-PROGRAMMES REQUIS

- Sous-programme principal :

  SIMOB (le programme principal est le programme conversationnel).

- Gestion des sous-programmes :

  SIMIL

- Lecture des paramètres, initialisations :

  LIPAR, LIMØD, V5INI, V6INI

- Allocation dynamique de mémoire :

  ALLØC, ALLØ2

- Lecture/transposition des données :

  LCDØR, LCDØI

- Calcul des indices :

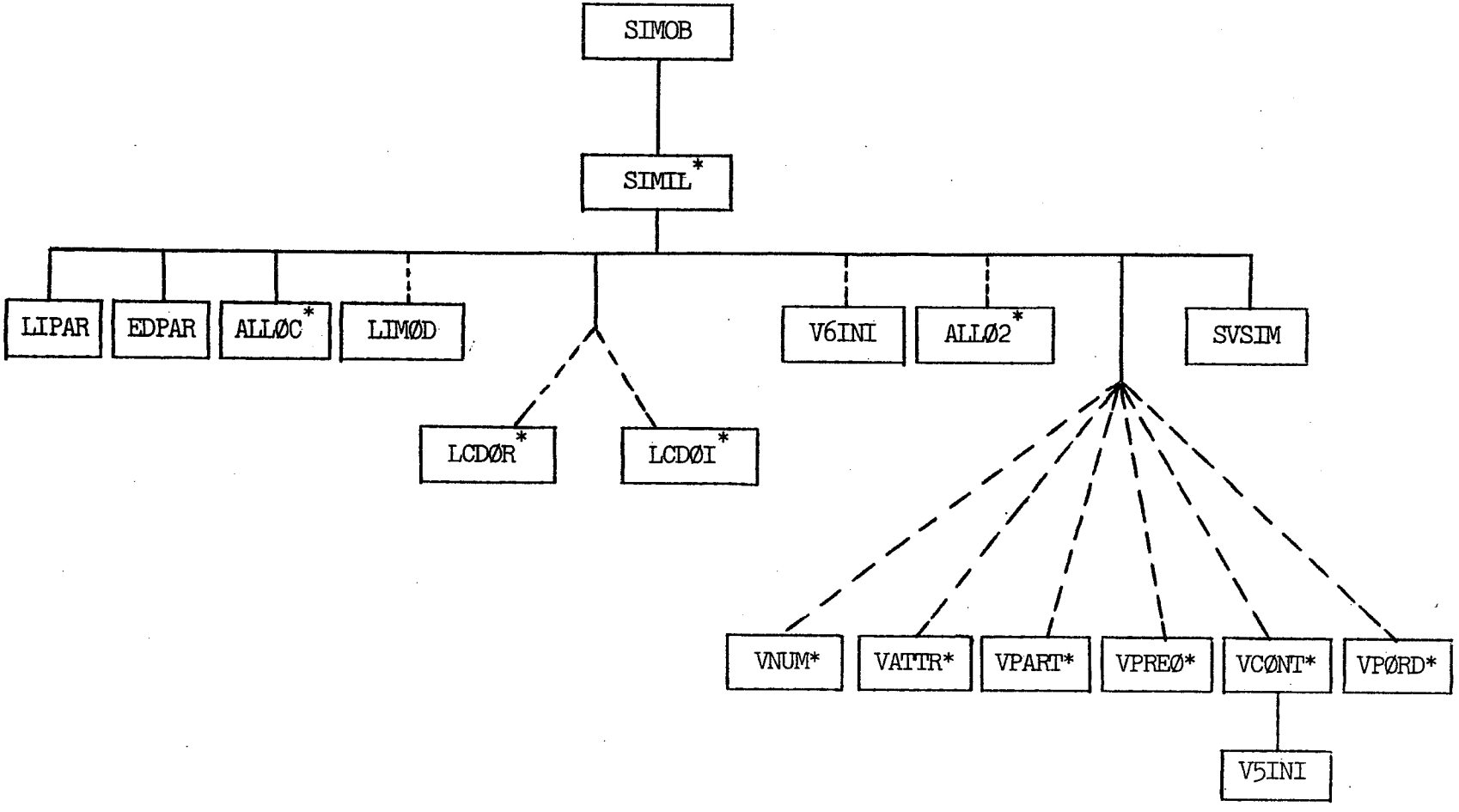
  VNUM, VATTR, VPART, VPREØ, VCØNT, VPØRD

- Impression des résultats :

  EDPAR, SVSIM

- Traitement des erreurs :

  ERRØR.



Arborescence des appels :

*: sous-programme susceptible d'appeler le sous-programme ERRØR.

4. FICHIERS UTILISES

* Fichier des paramètres (10) :
exploité en lecture par les sous-programmes LIPAR, LIMØD et V6INI ; fichier d'entrée avec format.

* Fichier de sortie (11) :
rempli par les sous-programmes EDPAR et ERRØR. Ce fichier contiendra un compte-rendu de l'exécution du programme : rappel des paramètres et éventuellement messages d'erreurs ; fichier de sortie avec format.

* Fichier des données (12) :
exploité en lecture par le sous-programme LCDØR (resp. LCDØI) pour des données réelles (resp. entières) ; fichier d'entrée avec format.

* Fichier des données (13) :
rempli et exploité par le sous-programme LCDØR (resp. LCDØI) pour des données réelles (resp. entières) ; fichier de travail sans format.

* Fichier transposé des données (14) :
rempli par LCDØR (ou LCDØI), exploité par le sous-programme de calcul des indices (VNUM, VATTR, VPART, VPREØ, VCØNT ou VPØRD) ; fichier de travail sans format.

* Fichier de sortie des indices (15) :
rempli par le sous-programme SVSIM. Ce fichier contiendra la matrice triangulaire inférieure des indices de proximité entre individus ; fichier de sortie avec format (5(2X,E14.7)).

5. TRANSMISSION ET VALIDATION DES DONNEES :

Les données doivent se trouver sur le fichier 12 où une ligne du fichier doit correspondre au codage d'un individu par la variable descriptive choisie ; le format de lecture figurant dans le fichier des paramètres (10).

Echanges d'informations sur supports magnétiques :

Sur fichier 13 : recopie des données ; écriture et exploitation par le sous-programme LCDØR (ou LCDØI).

Sur fichier 14 : fichier transposé des données ; rempli par LCDØR (ou LCDØI) ; exploité par le sous-programme de calcul d'indices approprié.

6. GESTION DES ERREURS

Chaque erreur détectée entraîne la modification d'une variable indicatrice d'erreur (NUMERR) initialisée à zéro et l'appel au sous-programme ERRØR. Ce sous-programme écrit un message sur le fichier (11) contenant le compte-rendu de l'exécution du programme puis rend la main au sous-programme appelant. La détection d'une erreur entraîne l'abandon du programme (toute les erreurs détectées sont fatales).

Erreurs détectées :

- dimension de tableau insuffisante (impression de la dimension nécessaire)
- indice de réduction inconnu
- division par zéro (variance nulle, toutes les mesures d'un individu nulles, ...).

ASPECTS INFORMATIQUES

DOSSIER DE PROGRAMMATION

SOUS-PROGRAMME SIMOB

1) Objet : en fait le programme principal dimensionnement des super-tableaux.
2) Description des paramètres :
aucun
3) Sous-programme requis :
SIMIL.

SOUS-PROGRAMME SIMIL

1) Objet : c'est en fait le sous-programme principal, il gère les différents autres sous-programmes.

2) Description des paramètres :

2-1. LGINTG INTEGER(E) : dimension du super-tableau des entiers.
2.2. LGREEL INTEGER(E) : dimension du super-tableau des réels.
2.3. INTG INTEGER(T) : super-tableau des entiers.
2.4. REEL REAL(T) : super-tableau des réels.
2.5. NUMERR INTEGER(ER/ES) : code des erreurs.

3) Sous-programmes requis :
LIPAR, EDPAR, ALLØC, LIMØD, LCDØR, LCDØI, V6INI, ALLØ2, VNUM, VATTR, VPART, VPREØ, VCØNT, VPØRD, SVSIM, ERRØR.

4) Sous-programme appelant :
SIMOB.

SOUS-PROGRAMME LIPAR :

1) Objet : lecture des paramètres.
2) Description des paramètres :

   2.1. IENT : INTEGER(E): numéro du fichier des paramètres à lire
   2.2. IPARAM : INTEGER(S): tableau des paramètres
   2.3. ITITRE : CHARACTER(S): titre de l'expérience
   2.4. IFMT : CHARACTER(S): format des données.

3) Sous-programme requis :
   aucun
4) Sous-programme appelant :
   SIMIL.

SOUS-PROGRAMME EDPAR :

1) Objet : édition d'une page rappelant les paramètres.
2) Description des paramètres :

   2.1. ISØRT : INTEGER(E) : numéro du fichier de sortie
   2.2. IPARAM : INTEGER(E) : tableau des paramètres
   2.3. ITITRE : CHARACTER(E) : titre de l'expérience
   2.4. IFMT : CHARACTER(E) : format des données

3) Sous-programme requis :
   aucun
4) Sous-programme appelant :
   SIMIL.

SOUS-PROGRAMME ALLØC :

1) Objet : simulation d'une allocation dynamique pour les tableaux nécessaires dans la suite du programme. Pour les tableaux d'entiers, on pioche dans INIG, les pointeurs de début commencent par I ; pour les réels, on pioche dans REEL, les pointeurs de début commencent par J. On vérifie également la place mémoire disponible.

2) Description des paramètres :

2.1. ISØRT : INTEGER(E): numéro du fichier de sortie
2.2. LGINTG: INTEGER(E): dimension du super-tableau des entiers
2.3. LGREEL: INTEGER(E): dimension du super-tableau des réels
2.4. NIND : INTEGER(E): nombre d'individus
2.5. NVAR : INTEGER(E): nombre de variables
2.6. NQ : INTEGER(E): dimension du tableau des indices de proximité
2.7. NST : INTEGER(E): nombre de sous-tableaux
2.8. NRGT : INTEGER(E): dimension du tableau des rangs
2.9. INDICE : INTEGER(E): type des variables
2.10. IIXT : INTEGER(S): pointeur du tableau IXT
2.11. IIYT : INTEGER(S): pointeur du tableau IYT
2.12. IIQT : INTEGER(S): pointeur du tableau IQT
2.13. NQT : INTEGER(S): dimension des tableaux de transposition des données
2.14. ILECT: INTEGER(S): pointeur du tableau LECT
2.15. INBMØD : INTEGER(S): pointeur du tableau NBMØD
2.16. INMØD : INTEGER(S): pointeur du tableau NMØD
2.17. ILDEBU: INTEGER(S): pointeur du tableau LDEBU
2.18. INRG : INTEGER(S): pointeur du tableau NRG
2.19. JXT : INTEGER(S): pointeur du tableau XT
2.20. JYT : INTEGER(S): pointeur du tableau YT
2.21. JQT : INTEGER(S): pointeur du tableau QT
2.22. JQ : INTEGER(S): pointeur du tableau Q
2.23. JWJ : INTEGER(S): pointeur du tableau WJ
2.24. JXLECT : INTEGER(S): pointeur du tableau XLECT
2.25. JPI : INTEGER(S): pointeur du tableau PI
2.26. JPIP : INTEGER(S): pointeur du tableau PIP
2.27. JSPI : INTEGER(S): pointeur du tableau SPI2
2.28. JSIGX: INTEGER(S): pointeur du tableau SIGX

2.29. JPPP : INTEGER(S): pointeur du tableau PPP

2.30. JPPH : INTEGER(S): pointeur du tableau PPH

2.31. JF2SP: INTEGER(S): pointeur du tableau F2SP

2.32. IFINTG : INTEGER(S): nombre total de places occupées par les tableaux d'entiers

2.33. IFREEL : INTEGER(S): nombre total de places occupées par les tableaux de réels

2.34. NUMERR : INTEGER(ER/ES): code des erreurs.

3) Sous-programme requis :
ERRØR

4) Sous-programme appelant :
SIMIL.

SOUS-PROGRAMME LCDØR :

1) Objet : lecture et transposition des données dans le cas de données réelles (cas des variables numériques).
Initialisation du tableau SIGX(*) qui pour chaque individu contient la racine carrée de la somme des carrés de ses mesures.

2) Description des paramètres :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers
2.2. IFMT : CHARACTER(E) : format des données
2.3. NIND : INTEGER(E) : nombre d'individus
2.4. NVAR : INTEGER(E) : nombre de variables
2.5. NQT : INTEGER(E) : dimension du tableau QT
2.6. SIGX : REEL(S) : tableau qui pour chaque individu contient la racine carrée de la somme des carrés de ses mesures.
2.7. QT : REEL(T) : tableau de travail pour la transposition des données
2.8. XT : REEL(T) : tableau de lecture des données
2.9. YT : REEL(T) : tableau d'écriture des données transposées
2.10. NUMERR: INTEGER(ER/ES) : code des erreurs

3) Sous-programme requis :
ERRØR
4) Sous-programme appelant :
SIMIL
5) Divers :
fichiers :
* fichier des données : IFDØN = IFICH(3)=12, fichier de lecture avec format
* fichier binaire des données : IFTAB = IFICH(4)=13 ; fichier de travail sans format
* fichier binaire des données transposées : IFTABT = IFICH(S)=14 ; fichier d'écriture sans format.

SOUS-PROGRAMME LCDØI :

1) Objet : lecture et transposition des données dans le cas de données entières. Initialisations de certains tableaux pour la suite du programme

2) Description des paramètres :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers
2.2. IFMT : CHARACTER(E) : format des données
2.3. NIND : INTEGER(E) : nombre d'individus
2.4. NVAR : INTEGER(E) : nombre de variables
2.5. NQT : INTEGER(E) : dimension du tableau IQT
2.6. INDICE : INTEGER(E) : type des variables
2.7. IRD : INTEGER(E) : procédé de réduction (cas des attributs)
2.8. NST : INTEGER(E) : nombre de sous-tableaux (cas des tableaux de contingence)
2.9. NBMØD : INTEGER(E) : tableau des nombres de modalités par variable (quand il y a lieu)
2.10. NMØD : INTEGER(E) : tableau des nombres de variables par sous-tableau (quand il y a lieu)
2.11. PI : REEL(S) : tableau à initialiser dans le cas des attributs (voir "divers").
2.12. PIP : REEL(S) : tableau à initialiser dans le cas de tableaux de contingence (voir "divers").
2.13. PPP : REEL(S) : tableau à initialiser dans le cas de tableaux de contingence (voir "divers").
2.14. SPI2 : REEL(S) : tableaux à initialiser dans le cas de tableaux de contingence (voir "divers").
2.15. IQT : INTEGER(T) : tableau de travail pour la transposition des données.
2.16. IXT : INTEGER(T) : tableau de lecture des données.
2.17. IYT : INTEGER(T) : tableau d'écriture des données transposées.
2.18. NUMERR: INTEGER(ER/ES) : code des erreurs.

3) Sous-programme requis :
ERRØR

4) Sous-programme appelant :
SIMIL.

5) Divers :

- fichiers :
  * fichier des données : IFDØN = IFICH(3)=12 ; fichier de lecture avec format
  * fichier binaire des données : IFTAB = IFICH(4)=13 ; fichier de travail sans format
  * fichier binaire des données transposées : IFTABT = IFICH(4)=14 ; fichier d'écriture sans format.

- tableaux initialisés :

  cas des attributs :

  PI(NIND) : pour chaque individu contient la proportion d'attributs possédés.

  cas des tableaux de contingence :

  PIP(NIND,NST) : pour chaque individu et pour chaque sous-tableau contient la somme des mesures de l'individu pour le sous-tableau (noté Ki. dans la partie théorique).

  PPP(NST) : pour chaque sous-tableau contient la somme des mesures de tous les individus pour le sous-tableau (noté K.. dans la partie théorique).

  SPI2(NST) : pour chaque sous-tableau contient la somme des carrés des mesures de tous les individus pour le sous-tableau.

SOUS-PROGRAMME LIMØD

1) Objet : lecture du nombre de modalités par variable (variables partitions, préordinales ou préordonnances) ou du nombre de variables par sous-tableau (tableau de contingence).

2) Description des paramètres :

2.1. IENT : INTEGER(E): numéro du fichier des paramètres

2.2. N : INTEGER(E): nombre de variables ou de sous-tableaux

2.3. MODALI : INTEGER(S) : tableau du nombre de modalités par variable ou du nombre de variables par sous-tableau.

2.4. MØDMX : INTEGER(S) : nombre maximum de modalités par variable.

3) Sous-programme requis :

aucun

4) Sous-programme appelant :

SIMIL.

SOUS-PROGRAMME ALLØ2 :

1) Objet : sous-programme complétant l'allocation dynamique dans le cas de variables partitions, préordinales ou préordonnances ; où on a besoin d'un tableau de dimension égale au nombre maximum de modalités par variable MODMX ; nombre inconnu au moment de l'allocation dynamique générale.

2) Description des paramètres

2.1. ISØRT : INTEGER(E) : numéro du fichier de sortie.
2.2. LGINTG : INTEGER(E) : dimension du super-tableau des entiers.
2.3. LGREEL : INTEGER(E) : dimension du super-tableau des réels.
2.4. MØDMX : INTEGER(E) : nombre maximum de modalitésd'une variable.
2.5. IFINTG : INTEGER(ES) : nombre de places occupées par les tableaux d'entiers.
2.6. IFREEL : INTEGER(ES) : nombre de places occupées par les tableaux de réels.
2.7. ILSYM : INTEGER(S) : début du tableau LSYM.
2.8. JXNV : INTEGER(S) : début du tableau XNV.
2.9. NUMERR : INTEGER(ER/ES) : code des erreurs.

3) Sous-programme requis :
ERRØR.
4) Sous-programme requis :
SIMIL.

SOUS-PROGRAMME V6INI :

1) Objet : lecture du tableau des rangs entre modalités et initialisations de tableaux facilitant son accès (variables préordonnances).

2) Description des paramètres

2.1. IENT : INTEGER(E) : numéro du fichier des paramètres
2.2. NIND : INTEGER(E) : nombre d'individus
2.3. NVAR : INTEGER(E) : nombre de variables
2.4. NLSYM: INTEGER(E) : dimension du tableau LSYM
2.5. NRGT : INTEGER(E) : dimension du tableau NRG des rangs
2.6. NBMØD: INTEGER(E) : tableau du nombre de modalités par variable
2.7. LDEBU: INTEGER(S) : tableau de repérage du début des rangs entre ses modalités dans le tableau NRG des rangs.
2.8. NRG : INTEGER(S) : tableau des rangs
2.9. LSYM : INTEGER(S) : tableau repérant le rang entre deux modalités d'une variable.

3) Sous-programme requis :
aucun

4) Sous-programme appelant :
SIMIL

5) Divers :

Accès au tableau des rangs :

Les rangs sont rangés variable par variable. Pour une variable v l'adresse A^v_{ij} du rang entre sa modalité i et sa modalité j (avec j>i) est :

$A^v_{ij} = deb_v - 1 + (j*(j-1))/2 + i$;

où deb_v est l'adresse du premier rang pour la variable v dans le tableau NRG.

Le tableau LDEBU(NVAR) contient pour chaque variable v $deb_v - 1$;
Le tableau LSYM(NLSYM) contient pour chaque j (j>0 et j>NLSYM) (j*(j-1)/2
L'adresse A^v_{ij} devient alors : A^v_{ij} = LDEBU(v)+LSYM(y)+i.

SOUS-PROGRAMME VNUM :

1) **Objet** : calcul des indices de proximité entre individus décrits par des variables numériques.

2) **Description des paramètres** :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers
2.2. NIND : INTEGER(E) : nombre d'individus
2.3. NVAR : INTEGER(E) : nombre de variables
2.4. NQ : INTEGER(E) : dimension du tableau Q
2.5. SIGX : REAL(E) : tableau qui pour chaque individu contient la racine carrée de la somme des carrés de ses mesures (initialisé dans LCDØR)
2.6. Q : REEL(S) : tableau de la matrice triangulaire inférieure des indices de proximités
2.7. WJ : REEL(T) : tableau qui pour chaque individu contient η_i^j au moment du calcul de la contribution de la variable j (notation de la partie théorique)
2.8. XLECT : REEL(T) : tableau de lecture du fichier transposé des données
2.9. NUMERR: INTEGER(ER/ES) : code des erreurs.

3) **Sous-programme requis** :
ERRØR

4) **Sous-programme appelant** :
SIMIL

5) **Divers** :

* fichier transposé des données : IFTABT=IFICH(5)=14 ; fichier de lecture sans format.
* algorithme :

pour chaque variable j :

* lecture de la ligne du fichier correspondante
* calcul pour chaque individu i de η_i^j dans WI(I)
* calcul de la moyenne et de la variance
* ajout du tableau Q de la contribution de la variable j aux indices de proximité entre individus.

SOUS-PROGRAMME VATTR :

1) Objet : calcul des indices de proximité entre individus décrits par des attributs.

2) Description des paramètres :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers

2.2. NIND : INTEGER(E) : nombre d'individus

2.3. NVAR : INTEGER(E) : nombre de variables

2.4. NQ : INTEGER(E) : dimension du tableau Q

2.5. IRD : INTEGER(E) : procédé de réduction

2.6. PI : REAL(E) : tableau contenant pour chaque individu la proportion d'attributs possédés

2.7. Q : REAL(S) : tableau de la matrice triangulaire inférieure des indices de proximité

2.8. LECT : INTEGER(T) : tableau de lecture des données

2.9. WJ : REAL(T) : tableau qui pour chaque individu i, contient $n^{j}_{ird-1,i}$ au moment du calcul de la contribution de l'attribut j (notation de la partie théorique).

2.10. NUMERR: INTEGER(ER/ES) : code des erreurs.

3) Sous-programme requis :

ERRØR

4) Sous-programme appelant :

SIMIL

5) Divers :

* Fichier transposé des données : IFTAB = IFICH(5)=14 ; fichier de lecture sans format.

* algorithme :

pour chaque variable j :

* lecture de la ligne du fichier correspondante
* calcul pour chaque individu i de $n^{j}_{ird-1,i}$ dans WJ(i)
* calcul de la moyenne et de la variance
* ajout au tableau Q de la contribution de la variable j aux indices de proximité entre individus.

SOUS-PROGRAMME VPART :

1) Objet : calcul des indices de proximité entre individus décrits par des variables partitions.
2) Description des paramètres :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers
2.2. NIND : INTEGER(E) : nombre d'individus
2.3. NVAR : INTEGER(E) : nombre de variables
2.4. NQ : INTEGER(E) : dimension du tableau Q
2.5. MØDMX : INTEGER(E) : dimension du tableau XNV
2.6. NBMØD : INTEGER(E) : tableau contenant le nombre de modalités de chaque variable
2.7. Q : REAL(S) : tableau de la matrice triangulaire inférieure des indices de proximité
2.8. LECT : INTEGER(T) : tableau de lecture des données
2.9. XNV : REAL(T) : tableau contenant les fréquences des modalités de la variable courante
2.10. NUMERR: INTEGER(ER/ES) : code des erreurs.

3) Sous-programme requis :
ERRØR
4) Sous-programme appelant :
SIMIL
5) Divers :
* fichier transposé des données = IFTABT = IFICH(5)=14 ; fichier de lecture sans format.
* algorithme :
pour chaque variable j :
* lecture de la ligne du fichier correspondante
* détermination du tableau XNV
* calcul de la moyenne et de la variance
* ajout au tableau Q de la contribution de la variable j aux indices de proximité entre individus.

SOUS-PROGRAMME VPREØ :

1) Objet : calcul des indices de proximité entre individus décrits par des variables préordinales.
2) Description des paramètres :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers
2.2. NIND : INTEGER(E) : nombre d'individus
2.3. NVAR : INTEGER(E) : nombre de variables
2.4. NQ : INTEGER(E) : dimension du tableau Q
2.5. MØDMX : INTEGER(E) : dimension du tableau XNV
2.6. NBMØD : INTEGER(E) : tableau contenant le nombre de modalités de chaque variable
2.7. Q : REAL(S) : tableau de la matrice triangulaire inférieure des indices de proximité
2.8. LECT : INTEGER(T) : tableau de lecture des données
2.9. XNV : REAL(T) : tableau contenant les fréquences des modalités de la variable courante
2.10. NUMERR: INTEGER(ER/ES) : code des erreurs.

3) Sous-programme requis :
ERRØR
4) Sous-programme appelant :
SIMIL
5) Divers :
* fichier transposé des données : IFTABT = IFICH(5)=14 ; fichier de lecture sans format.
* algorithme
pour chaque variable j :
* lecture de la ligne du fichier correspondante
* détermination du tableau XNV
* calcul de la moyenne et de la variance
* ajout au tableau Q de la contribution de la variable j aux indices de proximité entre individus.

SOUS-PROGRAMME VCONT :

1) Objet : calcul des indices de proximité dans le cas d'un tableau de contingence.

2) Description des paramètres :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers
2.2. NIND : INTEGER(E) : nombre d'individus (par abus de langage)
2.3. NVAR : INTEGER(E) : nombre total de variables (par abus de langage)
2.4. NST : INTEGER(E) : nombre de sous-tableaux
2.5. NQ : INTEGER(E) : dimension du tableau Q
2.6. PPP : REAL(E) : tableau des k.. pour chaque sous-tableau
2.7. NMØD : INTEGER(E) : nombre de variables par sous-tableau
2.8. PIP : REAL(ES) : tableau des $p_{i.}$
2.9. PPH : REAL(ES) : tableau des $p_{.j}$
2.10. SPI2 : REAL(ES) : tableau des $\sum_j p_{i.}^2$
2.11. Q : REAL(S) : tableau de la matrice triangulaire inférieure des indices de proximité
2.12. LECT : INTEGER(T) : tableau de lecture des données
2.13. XLECT : REAL(T) : tableau des f_{ij}
2.14. F2SP : REAL(T) : tableau des $\left(\sum_h \frac{(f_h^i - p.h)^2}{p.h}\right)^{1/2}$
2.15. NUMERR : INTEGER(ER/ES) : code des erreurs.

3) Sous-programmes requis :
V5INI - ERRØR

4) Sous-programme appelant :
SIMIL

5) Divers :
* fichier transposé des données : IFTABT=IFICH(5)=14 ; fichier de lecture sans format.
* algorithme :
pour chaque sous-tableau
pour chaque "variable" du sous-tableau
* lecture de la ligne des données
* calcul de la moyenne et de la variance
* ajout de la contribution de la "variable" au tableau Q des indices de proximité.

SOUS-PROGRAMME V5INI :

1) Objet : initialisation du tableau F2SP contenant $(\sum_h \frac{(f_h^i - p.h)^2}{p.h})^{1/2}$ (cas des tableaux de contingence).

2) Description des paramètres

2.1. IFTABT : INTEGER(E) : fichier transposé des données
2.2. NIND : INTEGER(E) : nombre d'individus
2.3. NVAR : INTEGER(E) : nombre total de variables
2.4. NST : INTEGER(E) : nombre de sous-tableaux
2.5. NMØD : INTEGER(E) : nombre de variables par sous-tableau
2.6. PPP : REAL(E) : tableau des k..
2.7. PIP : REAL(ES) : tableau des $p_{i.}$
2.8. PPH : REAL(ES) : tableau des $p_{.j}$
2.9. F2SP : REAL(ES) : tableau des $(\sum_h \frac{(f_h^i - p.h)^2}{p.h})^{1/2}$
2.10. LECT : INTEGER(T) : tableau de lecture des données
2.11. NUMERR : INTERGER(ER/ES) : code des erreurs.

3) Sous-programme requis :
aucun
4) Sous-programme appelant :
VCONT
5) Divers :
fichier transposé des données IFTABT=14
fichier de lecture sans format.

SOUS-PROGRAMME VPØRD :

1) Objet : calcul des indices de proximité entre individus décrits par des variables préordonnances.

2) Description des paramètres :

2.1. IFICH : INTEGER(E) : tableau des numéros de fichiers
2.2. NIND : INTEGER(E) : nombre d'individus
2.3. NVAR : INTEGER(E) : nombre de variables
2.4. NLSYM : INTEGER(E) : dimension du tableau LSYM
2.5. NQ : INTEGER(E) : dimension du tableau Q
2.6. NRGT : INTEGER(E) : dimension du tableau des rangs NRG
2.7. LDEBU : INTEGER(E) : tableau d'adressage
2.8. LSYM : INTEGER(E) : tableau d'adressage
2.9. NRG : INTEGER(E) : tableau des rangs
2.10. Q : REAL(S) : tableau de la matrice triangulaire inférieure des indices de proximités entre individus
2.11. LECT : INTEGER(T) : tableau de lecture des données
2.12. NUMERR: INTEGER(ER/ES) : code des erreurs.

3) Sous-programme requis :
ERRØR

4) Sous-programme appelant :
SIMIL

5) Divers :
* fichier transposé des données : IFTABT = IFICH(5)=14 ; fichier de lecture sans format.
* algorithme :
pour chaque variable j :
* lecture de la ligne du fichier correspondante
* calcul de la moyenne et de la variance
* ajout au tableau Q de la contribution de la variable j aux indices de proximité entre individus.

SOUS-PROGRAMME ERRØR :

1) Objet : traitement des erreurs, impression d'un message d'erreur sur le fichier ISØRT=IFICH(2)=11 ; fichier avec format.

2) Description des paramètres :

2.1. ISØRT : INTEGER(E) : numéro du fichier de sortie
2.2. NUMERR : INTEGER(E) : code de l'erreur
2.3. ICDERR : INTEGER(E) : dimension de tableau nécessaire en cas de place mémoire insuffisante ou individu ou variable responsable de l'erreur.

3) Sous-programme requis :
aucun

4) Sous-programme appelant :
SIMIL, ALLØC, ALLØ2, LCDØR, LCDØI, VNUM, VATTR, VPART, VPREØ, VCØNT et VPØRD.

SOUS-PROGRAMME SVSIM :

1) Objet : impression des indices de proximité sur le fichier IFSIM=IFICH(6)=15

2) Description des paramètres :

2.1. IFSIM : INTEGER(E) : numéro du fichier de sortie
2.2. NQ : INTEGER(E) : dimension du tableau Q
2.3. Q : REAL (E) : tableau de la matrice triangulaire inférieure des indices de proximité entre individus.

3) Sous-programme requis :
aucun

4) Sous-programme appelant :
SIMIL

5) Divers :

* format de sortie : (5(2X,E14.7))
* les indices dans le tableau Q de la matrice triangulaire inférieure sont rangés (et imprimés) de la manière imposée par les normes MØDULAD (ligne par ligne).

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique